involve

a journal of mathematics

msp

2011 vol. 4, no. 2

involve

pjm.math.berkeley.edu/involve

See inside back cover or http://pjm.math.berkeley.edu/involve for submission instructions.

The subscription price for 2011 is US $100/year for the electronic version, and $130/year (+$35 shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFlow™ from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

http://msp.org/

A NON-PROFIT CORPORATION

Typeset in LaTeX

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.103

The visual boundary of $\mathbb{Z}^2$

Kyle Kitzmiller and Matt Rathbun

(Communicated by Kenneth S. Berenhaut)

We introduce ideas from geometric group theory related to boundaries of groups. We consider the visual boundary of a free abelian group, and show that it is an uncountable set with the trivial topology.

1. Introduction

The study of a metric space can often be facilitated by considering it in the large scale, or by studying asymptotic phenomena. For instance, adding a boundary to compactify (or, more generally, "bordify") a metric space is a key tool in understanding the space and its isometry group. A classical example is the hyperbolic space $\mathbb{H}^n$, with its boundary sphere at infinity. Isometries of $\mathbb{H}^n$ extend to homeomorphisms of the boundary, and can be classified by their fixed points on the boundary. More generally, any Gromov hyperbolic space (that is, a space with large-scale negative curvature) has such a naturally defined boundary at infinity [Bridson and Haefliger 1999].

In trying to understand the geometry of groups, it is often useful to regard the group as a metric space by choosing a generating set, and forming the associated *Cayley graph*, which will be defined below. The metric induced by declaring all edges in the Cayley graph to have length one is called the *word metric* on the group. It would seem quite natural to define a boundary for groups directly from the word metric, and this works well if the group is Gromov hyperbolic. In general, however, there are obstructions to the usefulness of this boundary, as we will see below. This note explores properties of the visual boundary for groups, introducing the needed definitions along the way. The main result is that the visual boundary of $\mathbb{Z}^2$ (denoted $\partial_\infty(\mathbb{Z}^2)$) with the standard generating set possesses the trivial topology on an uncountable set. Indeed, there are many groups which have so called "quasi-flats", or quasi-isometric embeddings of $\mathbb{Z}^2$. We will see that the boundary of any such group will inherit the unpleasant properties of $\partial_\infty(\mathbb{Z}^2)$.

MSC2000: 20F05, 20F69, 51F99.
Keywords: boundary, visual boundary, Cayley graph, $\mathbb{Z}^2$, geodesic ray, quasi-isometry.
The authors were supported by VIGRE NSF grant no. 0636297.

The exposition is intended to be readable for a student who has had a first course in topology and metric spaces, and who is familiar with the definition and the most basic examples of groups. (We also mention the axiom of choice.) On the other hand, we hope that the paper will be a nontrivial read for working mathematicians in other areas.

2. Background

Metric notions. We review here some useful definitions from metric geometry.

Definition 2.1.

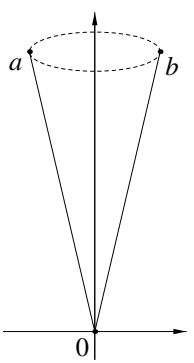
- A *geodesic segment*, *ray*, or *line* in a metric space X is an isometric embedding of $[0, a]$, $[0, \infty)$, or $\mathbb{R}$ into X. Thus, for instance, a geodesic line is a map $f : \mathbb{R} \to X$ such that $d_X(f(t_1), f(t_2)) = |t_1 - t_2|$ for all $t_1, t_2 \in \mathbb{R}$. We say a geodesic ray is *from* x_0 or *based at* x_0 if $f(0) = x_0$.
- A metric space is called a *geodesic space* if any two points in the space can be joined by a geodesic segment.
- Suppose (X, d) is a metric space, and $Y \subset X$ is connected. There are two natural ways to metrize Y. The *subspace metric* is $d_Y : Y \to \mathbb{R}_{\geq 0}$ defined by $d_Y(y_1, y_2) = d(y_1, y_2)$. Alternatively, the *path metric* is $d_{\text{path}} : Y \to \mathbb{R}_{\geq 0}$ defined by

$$d_{\text{path}}(y_1, y_2) = \inf\big\{\text{length}(\gamma) \mid \gamma \text{ is a path in } Y \text{ connecting } y_1 \text{ to } y_2\big\}.$$

- A geodesic space is called *(geodesically) complete* if every geodesic segment can be extended infinitely in both directions.
- A metric space is called *proper* if closed balls are compact. (This is needed for certain kinds of limiting arguments.)

Example 2.2. Consider the (half-)cone

$$X = \big\{(x, y, z) \mid x^2 + y^2 = \tfrac{1}{25}z^2,\ z \geq 0\big\},$$



a portion of which is shown in the figure. We claim that X is not geodesically complete, when considered with the path metric. Indeed, take a geodesic segment from the point $a = (-1, 0, 5)$ to the origin $(0, 0, 0)$; this coincides with a straight-line segment in space. Trying to extend this geodesic to $b = (1, 0, 5)$ presents a problem. The length of the two segments would be $2\sqrt{26}$, whereas the distance between the two points a and b is *at most* π, because we can go from one point to the other along the circle $\{z = 5\} \cap X$, which has radius 1. Certainly, if we try to extend the geodesic to any

other point on X, we will face the same difficulty: there is a shorter path "around" the cone, rather than going through the cone point.

Example 2.3. Let X be an infinite-dimensional Hilbert space. Then X is not proper, because the closed unit ball is not compact. To see this, take an orthonormal basis, $\{v_\alpha\}$ for X. Then any countably infinite sequence of the $\{v_\alpha\}$ is a sequence with no convergent subsequence, since the distance between any two elements is

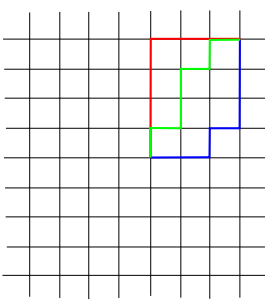
$$\|v_\alpha - v_\beta\| = \sqrt{\langle v_\alpha, v_\alpha\rangle + \langle v_\beta, v_\beta\rangle - \langle v_\alpha, v_\beta\rangle - \langle v_\beta, v_\alpha\rangle} = \sqrt{\|v_\alpha\| + \|v_\beta\|} = \sqrt{2}.$$

Cayley graphs. The construction of a Cayley graph is a central tool in geometric group theory, allowing us to associate a metric space to a group with a given presentation.

Definition 2.4. Let $G = \langle S \mid R\rangle$ be a group with generating set S and relations R. We define a graph $\mathrm{Cay}(G, S)$ whose vertices correspond to elements of G, and with edges between $g, h \in G$ if there exists $s \in S \cup S^{-1}$ so that $g = h \cdot s$. We give the resulting graph the graph metric, whereby each edge has length 1, and the distance between vertices is the length of the shortest path between them.

Remark 2.5. For any two elements $g, h \in G$, the distance from g to h in $\mathrm{Cay}(G, S)$ is just the length of the shortest word w in $S \cup S^{-1}$ such that $g = h \cdot w$.

Example 2.6. $\mathrm{Cay}(\mathbb{Z}^2, \{(1, 0), (0, 1)\})$ is just the integer grid. Consider a path from the origin to any other point (m, n) of $\mathbb{Z}^2$. This path consists of a union of horizontal and vertical segments between the integer coordinate points of the graph, the vertices (see figure). There are some crucial differences from familiar metric spaces like $\mathbb{R}^2$ with the Euclidean metric: there is more than one path of minimum length between the origin and (m, n) unless $m = 0$ or $n = 0$, and there is no unique prolongation of geodesic segments to rays.



The distance from (m, n) to (k, l) is $|m - k| + |n - l|$ (the ℓ^1 distance). Notice that $(m, n) = (k, l) \pm |m - k|(1, 0) \pm |n - l|(0, 1)$, so the distance is the length of the smallest word s composed of letters from $\{\pm(0, 1), \pm(1, 0)\}$ such that $(m, n) = (k, l) + s$.

Alternatively, one could consider embedding the integer grid into $\mathbb{R}^2$, and take the metric on $\mathbb{Z}^2$ to be the path metric induced by this inclusion.

Remark 2.7. This graph is not determined by a group, but clearly depends on the choice of a generating set S. To accommodate this, in the next section we introduce the notion of quasi-isometry.

Quasi-isometries. Often, we want to say that two spaces share some of the same large-scale geometric features, even when they are not isometric. To this end, we introduce the concept of *quasi-isometry*. This is like isometry, but allows for some bounded error in the form of a multiplicative and an additive factor. We will find that many notions about metric spaces can be "quasified".

Definition 2.8.

- We say a map between two metric spaces, $f : (X, d_X) \to (Y, d_Y)$ is a *quasi-isometric embedding* for some $k \geq 1$, $c \geq 0$, if for every $x_1, x_2 \in X$,

$$\frac{1}{k} d_X(x_1, x_2) - c \leq d_Y(f(x_1), f(x_2)) \leq k d_X(x_1, x_2) + c.$$

- We say that a quasi-isometric embedding, $f : (X, d_X) \to (Y, d_Y)$, is a *quasi-surjection* if there exists a $D > 0$ such that for every $y \in Y$, there is an $x \in X$ such that $d_Y(y, f(x)) < D$.

  If $f : (X, d_X) \to (Y, d_Y)$ is a quasi-isometric embedding which is also a quasi-surjection, then we say that f is a *quasi-isometry* and we say that (X, d_X) and (Y, d_Y) are *quasi-isometric*.

  In particular, a quasi-isometry admits a quasi-inverse. When we compose a quasi-isometry with a quasi-inverse, we almost get the identity. But, as with most things "quasi", we might be off by a multiplicative and additive constant.

- If $f : (X, d_X) \to (Y, d_Y)$ is a quasi-isometry, a *quasi-inverse* is a quasi-isometric embedding $g : (Y, d_Y) \to (X, d_X)$ so that for some $k \geq 1$, $c \geq 0$, for all $x_1, x_2 \in X$,

$$\frac{1}{k} d_X(x_1, x_2) - c \leq d_X(g \circ f(x_1), g \circ f(x_2)) \leq k d_X(x_1, x_2) + c,$$

  and for all $y_1, y_2 \in Y$,

$$\frac{1}{k} d_Y(y_1, y_2) - c \leq d_Y(f \circ g(y_1), f \circ g(y_2)) \leq k d_Y(y_1, y_2) + c.$$

Example 2.9. $\mathbb{R}$ is $(1, 1)$-quasi-isometric to $\mathbb{Z}$. Consider $f : \mathbb{R} \to \mathbb{Z}$, defined by $f(x) = \lfloor x \rfloor$, the floor function. Then for all $x, y \in \mathbb{R}$,

$$|x - y| - 1 \leq \big|\lfloor x \rfloor - \lfloor y \rfloor\big| \leq |x - y| + 1.$$

This map is clearly surjective.

Further, the inclusion $g : \mathbb{Z} \hookrightarrow \mathbb{R}$ is a quasi-inverse, since $f \circ g(n) = n$ for any $n \in \mathbb{Z}$ and $g \circ f(x) = \lfloor x \rfloor$ for any $x \in \mathbb{R}$. So, if $m, n \in \mathbb{Z}$,

$$|m - n| = |f \circ g(m) - f \circ g(n)| = |m - n|,$$

and if $x, y \in \mathbb{R}$,

$$|x-y|-1 \le |g\circ f(x) - g\circ f(y)| \le |x-y|+1.$$

Example 2.10. $\mathbb{R}^2$ is $(2,2)$-quasi-isometric to $\mathbb{Z}^2$. We will go through the calculation, but the idea is simple: rounding points in the plane down to points in the integer grid never distorts distances by too much, even when you change from ℓ^2 to ℓ^1 distance. Consider $f : \mathbb{R}^2 \to \mathbb{Z}^2$, defined by $f(x,y) = (\lfloor x\rfloor, \lfloor y\rfloor)$. Then, for any $(a,b), (x,y) \in \mathbb{R}^2$,

$$\begin{aligned} d_{\mathbb{Z}^2}(f(a,b), f(x,y)) &= \big|\lfloor x\rfloor - \lfloor a\rfloor\big| + \big|\lfloor y\rfloor - \lfloor b\rfloor\big| \\ &\le (|x-a|+1) + (|y-b|+1) \quad \text{(as above)} \\ &\le 2\max\{|x-a|, |y-b|\} + 2 \\ &\le 2\sqrt{(\max\{|x-a|, |y-b|\})^2} + 2 \\ &\le 2\sqrt{(x-a)^2 + (y-b)^2} + 2 \\ &= 2d_{\mathbb{R}^2}((a,b),(x,y)) + 2, \end{aligned}$$

and

$$\begin{aligned} d_{\mathbb{Z}^2}(f(a,b), f(x,y)) &= \big|\lfloor x\rfloor - \lfloor a\rfloor\big| + \big|\lfloor y\rfloor - \lfloor b\rfloor\big| \\ &\ge (|x-a|-1) + (|y-b|-1) \quad \text{(also as above)} \\ &\ge d_{\mathbb{R}^2}((a,b),(x,y)) - 2 \quad \text{(by the triangle inequality)} \\ &\ge \tfrac{1}{2} d_{\mathbb{R}^2}((a,b),(x,y)) - 2. \end{aligned}$$

It is easy to see that the inclusion $g : \mathbb{Z}^2 \hookrightarrow \mathbb{R}^2$ is a quasi-inverse; the composition $g \circ f : \mathbb{R}^2 \to \mathbb{R}^2$ moves points no more than $\sqrt{2}$.

Remark 2.11. Above, we used quasi-isometry constants $k = 2$, $c = 2$. It is a nice exercise to show that $k = \sqrt{2}$, $c = 2$ are actually the best constants possible. But often we will not care what the constants actually are — only that they exist.

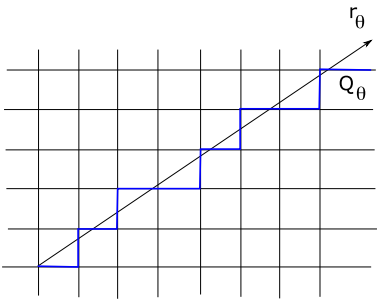
Definition 2.12. A *quasi-geodesic* is a quasi-isometric embedding of the real line into a space. That is, a map $f : \mathbb{R} \to X$ such that for some $k \ge 1$, $c \ge 0$, for all $t_1, t_2 \in \mathbb{R}$,

$$\frac{1}{k}|t_1 - t_2| - c \le d_X(f(t_1), f(t_2)) \le k|t_1 - t_2| + c.$$

Quasi-geodesics are useful, for instance, in discrete spaces: they can sit still for a bounded period of time, and can make jumps of bounded size, but in the large scale they proceed with distance roughly equal to time elapsed.

Example 2.13. Denote by r_θ the real ray in $\mathbb{R}^2$ from the origin that makes an angle of θ with the positive x-axis. Then we can consider the image of r_θ under the quasi-isometry $f : \mathbb{R}^2 \to \mathbb{Z}^2$ from Example 2.10. The result is a (disconnected) quasi-geodesic in $\mathrm{Cay}(\mathbb{Z}^2, \{(1, 0), (0, 1)\})$.

In this case, if we connect successive lattice points of $f \circ r_\theta$ with geodesic segments, the result is a geodesic ray in $\mathrm{Cay}(\mathbb{Z}^2, \{(1, 0), (0, 1)\})$, as in the figure. Call this ray Q_θ.



Next, as promised, we confirm that the word metric is independent of the choice of generating set, up to quasi-isometry.

Proposition 2.14. *If G is a finitely generated group with two (finite) generating sets S and S', then* $\mathrm{Cay}(G, S)$ *is quasi-isometric to* $\mathrm{Cay}(G, S')$.

Proof. The identity map will be shown to be a quasi-isometry. Say $|S| = k$, $|S'| = l$, let d_S be the distance function in $\mathrm{Cay}(G, S)$, and $d_{S'}$ in $\mathrm{Cay}(G, S')$. Then, since S and S' are finite, let $m = \max\{d_{S'}(s, e) \mid s \in S\}$, and $n = \max\{d_S(s', e) \mid s' \in S'\}$, where $e \in G$ is the identity element.

Then, every element $g \in G$ can be written as a word in S'. And each of those generators can be written as words of S, each of length at most n. So

$$d_S(g, e) \leq n \cdot d_{S'}(g, e).$$

To get the second inequality, the argument is reversed: $d_{S'}(g, e) \leq m \cdot d_S(g, e)$. So letting $k = \max\{m, n\}$ yields the quasi-isometry inequality.

The argument is completed by noting that for any $g, h \in G$,

$$d(g, h) = d(h^{-1}g, e). \qquad \square$$

Now we can speak unambiguously about the large-scale geometry of groups — those properties of groups that are invariant under quasi-isometry.

The visual boundary.

Notation 2.15. Let X be a geodesic space. For $x_0 \in X$, we define

$$\mathbb{G}_{x_0}(X) = \{\text{unit speed geodesic rays from } x_0\}.$$

We will suppress X from the notation and simply write $\mathbb{G}_{x_0}$.

We want to think of light traveling along geodesics in the space X. So we think of the visual boundary as the set of all points one can "see" at infinity, standing at the point x_0.

We give $\mathbb{G}_{x_0}$ the topology of uniform convergence on compact sets. Recall:

Definition 2.16. Let (X, d) be a metric space and Y a topological space. Given a fixed element $f \in X^Y = \{\text{functions } g : Y \to X\}$, a compact set K of Y and a number $\epsilon > 0$, we let

$$B_K(f, \epsilon) = \left\{ g \in X^Y \;\middle|\; d(f(y), g(y)) < \epsilon \text{ for all } y \in K \right\}.$$

The sets $B_K(f, \epsilon)$ form a basis for the *topology of uniform convergence on compact sets* on X^Y.

So $\mathbb{G}_{x_0} \subset X^{\mathbb{R}}$ inherits the subspace topology. Roughly, if the images of two rays are "close" on large compact sets, then the rays are "close". And a sequence of rays converges to a limiting ray if the rays of the sequence agree with the limit on larger and larger compact sets.

Sometimes, however, if we "look" in different directions, we see the same point at infinity. To make this precise:

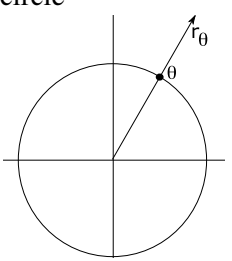
Definition 2.17. We say that two geodesic rays, g and f, are *asymptotic* if there exists an $M \in \mathbb{R}$ such that $d(f(t), g(t)) \leq M$ for all t. This is an equivalence relation on rays. We will write $f \sim g$, and denote the equivalence class of a ray $f \in \mathbb{G}_{x_0}$ by $[f]$, so $[f] = \{g \mid f \sim g\}$.

Definition 2.18. The *visual boundary* of a geodesic space X at a point x_0, denoted $\partial_\infty(X, x_0)$, is defined to be $\mathbb{G}_{x_0}/\sim$, with the quotient topology. Let $\pi_{x_0} : \mathbb{G}_{x_0} \to \partial_\infty(X, x_0)$ be the natural projection map.

Example 2.19. The visual boundary of $\mathbb{R}^2$ at $(0, 0)$ is homeomorphic to the unit circle, S^1.

Again, the idea is simple: every geodesic ray from the origin corresponds to exactly one point on the unit circle, and exactly one point at infinity.

Proof. Define a function $H : S^1 \to \partial_\infty(\mathbb{R}^2, (0, 0))$ by $H(\theta) = \pi(r_\theta)$, where r_θ is the straight line ray from the origin through the point on the unit circle corresponding to θ (see figure).



To show that this map is a bijection, note that given any two distinct points on the circle, θ and ϕ, the geodesic rays r_θ and r_ϕ diverge. That is, given any M, there exists some T such that $d(r_\theta(t), r_\phi(t)) > M$ for all $t > T$. Further, H is clearly surjective, as the only geodesic rays in $\mathbb{R}^2$ are straight line rays.

To show that the map is continuous, we will examine open balls about arbitrary points. Used implicitly in the remainder of the proof is the fact that H and π are bijections.

Assume V is open in $\partial_\infty(\mathbb{R}^2, (0,0))$. Then $H^{-1}(V) = \{r(1) \mid r \in \pi^{-1}(V)\}$. Now, consider an arbitrary point, $r^*(1) \in H^{-1}(V) \subset S^1$. We know what the basis of open sets in $\mathbb{G}$ looks like: it consists of the $B_K(f, \epsilon)$. So there exists an ϵ^* and a compact set $K = \{1\}$ such that the ball $B_{\{1\}}(r^*, \epsilon^*)$ is in $\pi^{-1}(V)$, because $\pi(r^*)$ is in V and $\pi^{-1}(V)$ is open. Then,

$$\begin{aligned} H^{-1}\big(\pi\big(B_{\{1\}}(r^*, \epsilon^*)\big)\big) &= H^{-1}\big(\pi\big(\{r \mid d(r(t), r^*(t)) < \epsilon^*, t \in \{1\}\}\big)\big) \\ &= H^{-1}\big(\pi\big(\{r \mid d(r(1), r^*(1)) < \epsilon^*\}\big)\big) \\ &= \{r(1) \mid d(r(1), r^*(1)) < \epsilon^*\} = B(r^*, \epsilon^*) \subset S^1. \end{aligned}$$

Now, assume W is open in S^1. We want to show that $H(W)$ is open. Consider any ray r^* such that $\pi(r^*) \in H(W)$. Then we know there exists an ϵ^* such that $B(r^*(1), \epsilon^*) = \{r(1) \mid d(r(1), r^*(1)) < \epsilon^*\} \subset W$. Then,

$$\begin{aligned} H(B(r^*(1), \epsilon^*)) &= \big\{\pi(r) \mid d(r(1), r^*(1)) < \epsilon^*\big\} \\ &= \big\{\pi(r) \mid d(r(t), r^*(t)) < \epsilon^*, t \in \{1\}\big\} = \pi(B_{\{1\}}(r^*, \epsilon^*)). \end{aligned}$$

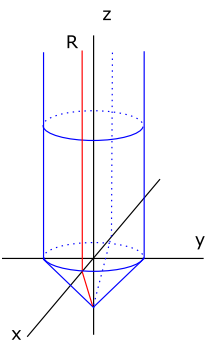
Since, in this case, $\pi^{-1}(\pi(B_{\{1\}}(r^*, \epsilon^*))) = B_{\{1\}}(r^*, \epsilon^*)$ is open, so is its image. Thus, given any point $\pi(r^*)$ in $H(W)$, there is an open set around this point contained in $H(W)$. We conclude that $H(W)$ is open, and ultimately that H is a homeomorphism between $\partial_\infty(\mathbb{R}^2, (0,0))$ and S^1. □

We would like a way to talk about the visual boundary of a metric space, without reference to a specified basepoint. Unfortunately, there are many cases when the visual boundary changes if we use a different basepoint.

Example 2.20 [Papadopoulos 2005]. Consider the set

$$R = \{(x, y, z) \mid x = 1, y = 0, z \geq 0\} \cup \{(x, y, z) \mid y = 0, z = 5x - 5, 0 \leq x \leq 1\}.$$

Rotate it around the z-axis to obtain a pencil-shaped surface X, considered with the path metric. If we take our basepoint to be $(0, 0, -5)$, then R is a geodesic ray from the basepoint, as is any rotation of R about the z-axis. So $\mathbb{G}_{(0,0,-5)}$ is a circle's worth of rays. If we take our basepoint to be $(1, 0, 0)$, on the other hand, the only geodesic ray from the basepoint is the ray $\{(x, y, z) \mid x = 1, y = 0, z \geq 0\}$. So $\mathbb{G}_{(1,0,0)}$ has a single ray.



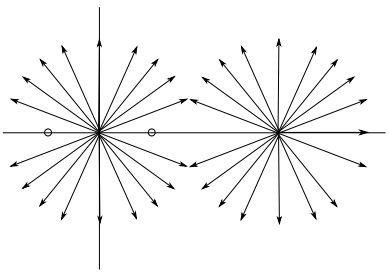
Notice, however, that all the rays in $\mathbb{G}_{(0,0,-5)}$ are asymptotic, since the distance between any two is bounded by π (in the path

metric). So, when we take the quotient, we get

$$\partial_\infty(X, (0, 0, -5)) \cong \partial_\infty(X, (1, 0, 0)) \cong \{\text{point}\}.$$

In this example, $\mathbb{G}$ depended on choice of basepoint, but the topological space $\partial_\infty(X)$ did not. In some spaces, however, even the visual boundary will change with the basepoint.

Example 2.21. Consider $X = \mathbb{R}^2 \setminus \{(1, 0), (-1, 0)\}$. Then if we choose the basepoint $(0, 0)$, there is a geodesic ray in every direction except along the positive and negative x-axes. So $\partial_\infty(X, (0, 0)) \cong (0, \pi) \cup (\pi, 2\pi)$. However, if we choose the basepoint $(3, 0)$, there is a geodesic ray in every direction except towards the negative x-axis. So $\partial_\infty(X, (2, 0)) \cong (-\pi, \pi)$ (see figure). This shows that the visual boundary of a twice-punctured plane depends on the choice of basepoint.



Fortunately, all is not lost.

Proposition 2.22. *Given two points x_1 and x_2 in a geodesic space X, let $L : X \to X$ be an isometry carrying x_1 to x_2. Then $\partial_\infty(X, x_1)$ is homeomorphic to $\partial_\infty(X, x_2)$.*

Proof. Isometries preserve geodesicity, so $\mathbb{G}_{x_1} \cong \mathbb{G}_{L(x_1)} = \mathbb{G}_{x_2}$. Further, the distance between geodesic rays is preserved, so $(\mathbb{G}_{x_1}/\sim) \cong (\mathbb{G}_{x_2}/\sim)$. □

Remark 2.23. When the isometry group of a space acts transitively on the space, as in the case of $\mathbb{R}^2$ or $\mathbb{Z}^2$, we can suppress the basepoint. So we will denote $\partial_\infty(X, x_0)$ as simply $\partial_\infty(X)$, π_{x_0} as π, and $\mathbb{G}_{x_0}$ as $\mathbb{G}$, when convenient.

Example 2.24. In light of Remark 2.23, $\partial_\infty(\mathbb{R}^2) \cong S^1$.

3. The case of $\mathbb{Z}^2$

Geodesic rays. We will henceforth abuse notation, and identify $\mathbb{Z}^2$ with its Cayley graph with respect to the standard generating set, $\text{Cay}(\mathbb{Z}^2, \{(1, 0), (0, 1)\})$, the integer grid (Example 2.6). We will also implicitly assume the basepoint to be $(0, 0)$. Geodesic paths consist of horizontal and vertical segments with no "backtracking". As noted above, geodesics are not unique. For example, there are twenty geodesic paths between $(0, 0)$ and $(3, 3)$, all of length 6 (see figure in Example 2.6).

It is clear, then, that for any ray f, the equivalence class $[f]$ is "large": there are many geodesics g such that $d(f, g) < M$ for all t.

Notation 3.1. An infinite ray in $\mathbb{Z}^2$, consisting of vertical and horizontal segments, can be expressed as an infinite string of the digits corresponding to each segment. Let 0, 1, 2, 3 and 4 represent east, north, west, south, and east respectively. Then any infinite ray in $\mathbb{Z}^2$ can be written as an infinite string over the alphabet $\{0, 1, 2, 3, 4\}$. (The redundant use of 0 and 4 for the eastward direction is to simplify later notation.)

If a ray is in the first quadrant, it can be written as a string over $\{0, 1\}$; in the second, $\{1, 2\}$; in the third, $\{2, 3\}$; and in the fourth, $\{3, 4\}$. To eliminate the only ambiguity, we adopt the convention that the east-pointing ray will be represented as the string $(\bar{0}) = (0, 0, 0, \dots)$ of all zeros. Given a geodesic ray $f \in \mathbb{Z}^2$, we will denote this expansion by $f = (f_1, f_2, f_3, \dots)$. Then if $m(f) = \min_n\{f_n\}$, we have $f_n \in \{m, m + 1\}$ for all n.

The topology on $\partial_\infty(\mathbb{Z}^2)$.

Definition 3.2. We will say a ray g in $\mathbb{Z}^2$ has *slope* θ if $g \sim Q_\theta$, where Q_θ is the image of the ray r_θ in $\mathbb{R}^2$ under the quasi-isometry in Example 2.10.

Note that not every ray has a slope. However, a ray cannot have more than one slope, because $\sim$ is transitive.

This sets us up to show that the visual boundary of $\mathbb{Z}^2$ is uncountable.

Proposition 3.3. $|\partial_\infty(\mathbb{Z}^2)| = \mathfrak{c}$, *the cardinality of the continuum.*

In order to prove this, we will describe an injection from S^1 into $\partial_\infty(\mathbb{Z}^2)$, and an injection from $\partial_\infty(\mathbb{Z}^2)$ into $[0, 4)$, making use of the quinary expansions described in the previous section.

Proof. The proof will proceed in two parts, exhibiting the two injections.

First, define the map $I : S^1 \to \partial_\infty(\mathbb{Z}^2)$ to be given by $I(\theta) = \pi(Q_\theta)$, where Q_θ is the quasi-isometric embedding of r_θ, the ray that passes through the point θ on the unit circle in $\mathbb{R}^2$. Then given any distinct $\theta, \phi \in S^1$, we have already seen that $\pi(Q_\theta) \neq \pi(Q_\phi)$. Thus I is an injection, and $\mathfrak{c} \leq |\partial_\infty(\mathbb{Z}^2)|$.

For the second injection, recall that any geodesic ray can travel in at most two directions. Hence, each ray corresponds to an infinite binary expansion. Let these binary strings be mapped to the interval $[0, 4)$ in the following way:

Let $B : \{0, 1\}^{\mathbb{N}} \to [0, 1]$ be the standard map from a binary expansion to the real number it represents. So

$$B((\epsilon_1, \epsilon_2, \epsilon_3, \dots)) = \sum_{n=1}^{\infty} \frac{\epsilon_n}{2^n}, \quad \text{where } \epsilon_n \in \{0, 1\} \text{ for all } n.$$

Now, for a quinary expansion,

$$(f_1, f_2, f_3, \dots) \in \{0, 1, 2, 3, 4\}^{\mathbb{N}},$$

let $m(f) = \min_n\{f_n\}$ as before. Then define a map $N : \mathbb{G}(\mathbb{Z}^2) \to [0, 4)$ by

$$N((f_1, f_2, f_3, \dots)) = m + B((f_1 - m, f_2 - m, f_3 - m, \dots)).$$

So for instance, $N(\bar{0}) = 0$ and

$$N((2, 3, 2, 3, 2, 3, \dots)) = 2 + B((0, 1, 0, 1, 0, 1, \dots)) = 2 + \tfrac{1}{3} = \tfrac{7}{3}.$$

It is easy to see (by uniqueness of binary expansions for the fractional part) that this map is injective from $\mathbb{G}(\mathbb{Z}^2) \to [0, 4)$. Thus $|\mathbb{G}(\mathbb{Z}^2)| \leq \mathfrak{c}$, so $|\partial_\infty(\mathbb{Z}^2)| \leq \mathfrak{c}$.

It follows that $|\partial_\infty(\mathbb{Z}^2)| = \mathfrak{c}$. □

Proposition 3.4. $\partial_\infty(\mathbb{Z}^2)$ *possesses the trivial topology.*

In other words, the only open sets in the visual boundary are the entire set and the empty set.

Proof. By the quotient topology on $\mathbb{G}/\sim$, a set $U \subset \partial_\infty(\mathbb{Z}^2)$ is open exactly when its preimage $\pi^{-1}(U)$ is open in $\mathbb{G}$. Assume that U is some nonempty open set in $\partial_\infty(\mathbb{Z}^2)$. Then, $W = \pi^{-1}(U)$ is also open and nonempty. We wish to show that U is the entire set. It suffices to show that given any $g \in \mathbb{G}$, $\pi(g) \in U$.

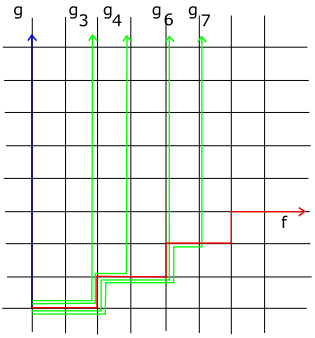
As W is open and nonempty, there is some geodesic ray f in W. Consider any ray g such that $m(g) = m(f)$. (This means that g and f are in the same quadrant.) We will show that given any compact set $K \subset [0, \infty)$ and $\epsilon > 0$, g has some representative $g_s \in [g]$ such that $g_s \in B_K(f, \epsilon) \subset W$. It will follow that $\pi(g) \in U$.

Let the compact set $K = [a, b]$ and $\epsilon > 0$ be given, and let $s = \lceil b \rceil \in \mathbb{Z}$. Then define the representative g_s of g as follows:

$$g_s(t) = \begin{cases} f(t) & \text{for } t \leq s, \\ f(s) + g(t) - g(s) & \text{for } t > s, \end{cases}$$

where the sum is group addition on $\mathbb{Z}^2$.

To clarify, consider the infinite binary expansion of g_s. It is identical to that of f for the first s steps, so $d(g_s(t), f(t)) = 0$ for $t \leq s$; afterwards it is identical to that of g, so $g_s \sim g$. This gives a sequence of rays asymptotic to g, in a neighborhood of f:



Clearly, $g_s \in B_K(f, \epsilon)$, so $\pi(g_s) \in \pi(B_K(f, \epsilon))$. Then since $B_K(f, \epsilon) \subset W$, $\pi(g_s) \in U$. Finally, since $g_s \sim g$, $\pi(g_s) = \pi(g)$. We conclude that $\pi(g) \in U$ and $g \in W$.

Recall that f and g are in the same quadrant because $m(f) = m(g) = m$. In particular, we see that the axis geodesic $h = (\overline{m+1}) \in B_K(f, \epsilon)$, where we take addition modulo 4.

We now take advantage of the fact that W is open. As $h \in W$, there must exist some ϵ' such that $B_K(h, \epsilon') \subset W$. Then let $j = (\overline{m+2})$ and let j_s be the representative function as above, so that $j_s \in B_K(h, \epsilon')$. Therefore $\pi(j) \in U$. Consequently, all axis directions are in U. By the same argument, then, we include in the set U the images of all other nonaxis geodesic rays g for which $m(g) \neq m(f)$. We can then conclude that given any geodesic ray $g \in \mathbb{Z}^2$, $\pi(g) \in U$.

By assuming only that U was open and nonempty, we showed that U contains all elements of $\partial_\infty(\mathbb{Z}^2)$. We conclude that $\partial_\infty(\mathbb{Z}^2)$ has the trivial topology. □

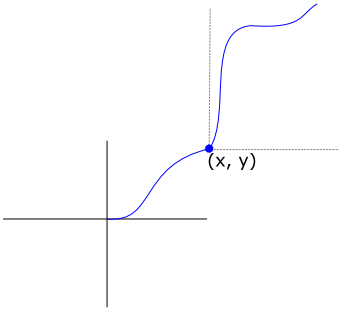
4. Further comments

Informally speaking, if we "zoom out" from $\mathbb{Z}^2$ by rescaling distances to be smaller and smaller, we limit to $\mathbb{R}^2$ with the ℓ^1-norm. (Formally, this construction is called the *asymptotic cone*, and $\mathrm{Cone}(\mathbb{Z}^2) = (\mathbb{R}^2, \ell^1)$.) We expect the same method of proof from above to show that the visual boundary of $(\mathbb{R}^2, \ell^1)$ is an uncountable set with the trivial topology. And in fact, this is true.

Proposition 4.1. $\partial_\infty((\mathbb{R}^2, \ell^1))$ *has the cardinality* $\mathfrak{c}$, *and the trivial topology.*

Proof. Geodesic rays are no longer restricted to vertical and horizontal segments, but they have a similar property. Let us first discuss geodesic rays that enter the interior of the first quadrant. Let $f(t) = (f_1(t), f_2(t))$ be a geodesic ray from the origin, passing through the point $f(t_0) = (x, y)$, with $x, y > 0$. Then for all $t > t_0$, $f_1(t) \geq x$, and $f_2(t) \geq y$. In other words, once a geodesic begins to move in a north-westerly direction, it can never again move toward the south or east (see

figure). A similar property, of course, also holds in the other quadrants.



There are more geodesic rays in this space than in $\mathbb{Z}^2$. But after we take the quotient, we get the same boundary. We will appeal to the Axiom of Choice. Certainly, $|\partial_\infty((\mathbb{R}^2, \ell^1))|$ is at least $\mathfrak{c}$, since each geodesic ray in $\mathbb{Z}^2$ includes as a geodesic ray into $(\mathbb{R}^2, \ell^1)$. Now, for each equivalence class of asymptotic rays $[f] \in \partial_\infty((\mathbb{R}^2, \ell^1))$, choose a representative geodesic ray, f. Then, as before, consider the image of this ray under the quasi-isometry from $\mathbb{R}^2$ onto $\mathbb{Z}^2$, and connect vertices by horizontal and vertical segments to get Q_f, a geodesic ray in $\mathbb{Z}^2$. Identifying Q_f and Q_g with their images in $\mathbb{R}^2$ by inclusion, we see that $f \sim Q_f$ and $g \sim Q_g$, so the map from $\partial_\infty((\mathbb{R}^2, \ell^1))$ to $\partial_\infty(\mathbb{Z}^2)$ is an injection. This establishes that $|\partial_\infty((\mathbb{R}^2, \ell^1))| = \mathfrak{c}$.

Next, we use an identical construction to the one above to show that the topology is trivial.

Let f, g be any arbitrary geodesic rays in the closure of quadrant I. We will show that given any compact $K \subset [0, \infty)$ and $\epsilon > 0$, g has a representative $g_b \in [g]$ such that $g_b \in B_K(f, \epsilon)$.

Let the compact set be $K = [a, b]$ and $\epsilon > 0$ be given. Then define the representative g_b of g as follows:

$$g_b(t) = \begin{cases} f(t) & \text{for } t \leq b, \\ f(b) + g(t) - g(b) & \text{for } t > b. \end{cases}$$

where now the sum is component addition on $\mathbb{R}^2$.

Just as before, this argument establishes that any open set containing a single ray in quadrant I contains all rays in quadrant I, and can be extended to show that any nonempty open set contains every ray. □

What's wrong with this state of affairs? This boundary completely fails to be Hausdorff: we can't separate any two directions at infinity. Convergence to a particular point in the boundary is meaningless.

To see some of the consequences of this finding, consider that a large class of groups have an undistorted free abelian subgroup; that is, a quasi-isometric

embedding of $\mathbb{Z}^2 \cong \langle a, b \rangle$, called a *quasi-flat*. These arise whenever two elements commute and there is no "shortcut" to words in those elements coming from a relator. Besides the obvious extension of the same argument to $\mathbb{Z}^n$, quasi-flats can also be found in right-angled Artin groups, as well as mapping class groups of surfaces. Papasoglu [1998] shows that every semi-hyperbolic group which is not hyperbolic contains such a quasi-flat. This includes fundamental groups of compact manifolds of nonpositive curvature. So this note shows, in particular, that any metric space containing a quasi-flat will have a bad boundary.

Acknowledgments

We would like to thank Moon Duchin for the inspiration of this project, as well as the for all of her motivation and support.

References

[Bridson and Haefliger 1999] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Math. Wissenschaften **319**, Springer, Berlin, 1999. MR 2000k:53038 Zbl 0988.53001

[Papadopoulos 2005] A. Papadopoulos, *Metric spaces, convexity and nonpositive curvature*, IRMA Lectures in Mathematics and Theoretical Physics **6**, European Mathematical Society (EMS), Zürich, 2005. MR 2005k:53042 Zbl 1115.53002

[Papasoglu 1998] P. Papasoglu, "Quasi-flats in semihyperbolic groups", *Proc. Amer. Math. Soc.* **126**:5 (1998), 1267–1273. MR 98j:20043 Zbl 0901.20024

Received: 2009-04-28 Revised: 2011-03-06 Accepted: 2011-03-07

kitz2103@gmail.com *Department of Mechanical Engineering, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-1323, United States*

m.rathbun@imperial.ac.uk *Department of Mathematics, University of California, One Shields Ave., Davis, CA 95616, United States*

Current address: *Department of Mathematics, Huxley Building 6M33, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom*
http://www2.imperial.ac.uk/~mrathbun

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.117

An observation on generating functions with an application to a sum of secant powers

Jeffrey Mudrock

(Communicated by Nigel Boston)

Suppose that $P(x)$, $Q(x) \in \mathbb{Z}[x]$ are two relatively prime polynomials, and that $P(x)/Q(x) = \sum_{n=0}^{\infty} a_n x^n$ has the property that $a_n \in \mathbb{Z}$ for all n. We show that if $Q(1/\alpha) = 0$, then α is an algebraic integer. Then, we show that this result can be used to provide a solution to Problem 11213(b) of the *American Mathematical Monthly* (2006).

1. Introduction and statement of results

This paper has two goals. One is to prove this general observation:

Theorem 1. *Suppose $P(x)$, $Q(x) \in \mathbb{Z}[x]$ are relatively prime polynomials with integer coefficients and their quotient is the generating function of an integer series*:

$$\frac{P(x)}{Q(x)} = \sum_{n=0}^{\infty} a_n x^n, \quad \text{with } a_n \in \mathbb{Z} \text{ for all } n.$$

Then the inverse of any root of Q is an algebraic integer.

The second goal is to apply this result to solve a problem from the *American Mathematical Monthly*:

Problem 11213 [AMM 2006]. *Proposed by Stanley Rabinowitz, Chelmsford, MA.* For positive integers n and m with n odd and greater than 1, let

$$S(n, m) = \sum_{k=1}^{(n-1)/2} \sec^{2m}\left(\frac{k\pi}{n+1}\right).$$

(a) Show that if n is one less than a power of 2, then $S(n, m)$ is a positive integer.

(b*) Show that if n does not have the form of Part (a), then there exists a positive integer m such that $S(n, m)$ is not an integer.

MSC2000: primary 11R04; secondary 11R18.
Keywords: algebraic number theory, generating functions, secant function.

The * indicates that no solution was known to the *Monthly* editors. (A solution to (a) was provided in [AMM 2008].) We solve part (b) of Problem 11213 by proving the contrapositive:

Theorem 2. *Let $n > 1$ be an odd integer. If, for every positive integer m, the sum*

$$S(n, m) = \sum_{k=1}^{(n-1)/2} \sec^{2m} \frac{k\pi}{n+1}$$

has an integer value, then $n + 1$ is a power of 2.

A similar result to Theorem 1 (but less general) had appeared before in the *Monthly*, as a problem proposed and solved by Michael Larsen:

Problem E 2993 [AMM 1983; 1986]. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ a complex numbers such that $\sum_1^n \alpha_i^m$ is an integer for every positive m; then the polynomial $\prod_1^n (x - \alpha_i)$ has integer coefficients.

Here is an outline of the paper. After recalling the necessary concepts from algebraic number theory in Section 2, we prove in Section 3 two intermediate results: $S(n, m)$ is always rational, and the generating function of the sequence $\{S(n, m)\}_{m>0}$ (for fixed odd $n > 0$) has integer coefficients. In Section 4 we prove Theorem 1, from which Theorem 2 follows easily given the intermediate results.

2. Background

We review some basic algebraic number theory, which is carefully laid out in [Stewart and Tall 2002], for example. (This citation will be abbreviated as [ST].)

An *algebraic number* is any zero of a polynomial with integer coefficients. An *algebraic integer* is any zero of a monic polynomial with integer coefficients. The set of algebraic numbers is a field, and the set of algebraic integers forms a ring [ST, Theorems 2.1 and 2.9].

For example, if p is prime, $\zeta_p = e^{2\pi i/p}$ is an algebraic integer since it is a zero of the polynomial $x^p - 1$.

The *minimal polynomial* of an algebraic number α is the monic polynomial $p(x)$ with rational coefficients and the smallest possible degree such that $p(\alpha) = 0$. All polynomials of which α is a root are divisible by p. For example, $r(x) = x^{p-1} + x^{p-2} + \cdots + x + 1 = (x^p - 1)/(x - 1)$ is the minimal polynomial of ζ_p.

Definition. If K is a field contained in L, we say that L is a field extension of K, and we denote this by $L : K$.

If K is a field and α is an algebraic number let $K(\alpha)$ denote the smallest field containing all the elements of K and α. One way to think about field extensions is that if $L : K$ is a field extension, then L has a natural structure as a vector space over

K. The dimension of this vector space, which is called the *degree*, is represented with $[L : K]$. If $[L : K]$ is a number the field extension is called finite. If H, K, and L are fields such that K is a subset of L and H is a subset of K, then

$$[L : H] = [L : K][K : H] \tag{1}$$

[ST, Theorem 1.10].

In algebraic number theory field extensions of the form $\mathbb{Q}(\alpha)$ are of interest. If α is an algebraic number, then $[\mathbb{Q}(\alpha) : \mathbb{Q}]$ equals the degree of the minimal polynomial of α [ST, Theorem 1.1]. A field K is called an *algebraic number field* if $[K : \mathbb{Q}]$ is finite. If $K = \mathbb{Q}(\alpha)$ and α is an algebraic number, then the ring of algebraic integers in K is finitely generated as an abelian group [ST, Theorem 2.16].

Definition. If $K = \mathbb{Q}(\alpha)$ is an algebraic number field of degree n, then there are n distinct monomorphisms $\sigma_1, \ldots, \sigma_n$ from K to $\mathbb{C}$. The *conjugates* of an element $\beta \in K$ are the numbers $\sigma_i(\beta)$ for all i between 1 and n.

The conjugates of an algebraic number α are the zeros of the minimal polynomial of α. For example, if $\alpha = \zeta_n = e^{2\pi i/n}$, where $n > 0$ is an integer, then α has $\phi(n)$ conjugates in $\mathbb{Q}(\alpha)$, where ϕ is the Möbius function. The conjugates of ζ_n are all the elements in the set

$$\{e^{2\pi ik/n} : (k, n) = 1\}.$$

This information can be found in [Milne 2009, page 93].

Definition. Let $K = \mathbb{Q}(\alpha)$ be an algebraic number field, and consider $\beta \in K$. The *trace* of β in K, denoted by $\mathrm{Tr}_K \beta$, is the sum of all the conjugates of β. The *norm* of β in K, denoted by $N_K(\beta)$, is the product of all of the conjugates of β.

Thus $\mathrm{Tr}_K \zeta_p = -1$ and $N_K(\zeta_p) = (-1)^{p-1}$ for p prime, where $K = \mathbb{Q}(\zeta_p)$. If one notes that

$$\frac{\zeta_n + \zeta_n^{-1}}{2} = \cos\frac{2\pi}{n}$$

and applies (1) one can see that the conjugates of $\alpha = \cos\dfrac{2\pi}{n}$ in $\mathbb{Q}(\alpha)$ are all the elements in the set

$$\left\{\cos\frac{2\pi k}{n} : (k, n) = 1,\ 0 < k < n/2\right\}. \tag{2}$$

A formal proof of this can be found in [Milne 2009, pages 95–96]. Also, as a consequence of Theorem 2.6(a), Lemma 2.13, and Lemma 1.7 of [ST], if α is an algebraic number its trace is rational; and as a consequence of Lemma 2.14 of the same reference, if α is an algebraic integer its norm is an integer.

3. Intermediate results

Lemma 3. *If $n > 1$ is odd and $m \geq 1$, the sum $S(n, m)$ of Theorem 2 is a rational number.*

Proof. We make use of the trigonometric identity $\sec^2 x = \dfrac{2}{\cos(2x)+1}$ to write $\sec^{2m} x = f(\cos 2x)$, where

$$f(x) := \left(\frac{2}{x+1}\right)^m.$$

Then, dropping m from the notation and introducing $N = n + 1$ for convenience, we can rewrite our sum as

$$\sum_{0<k<N/2} s(k), \quad \text{where } s(k) := f\left(\cos\frac{2\pi k}{N}\right). \tag{3}$$

We assume at first that $N/2$ is an odd prime. All the $s(k)$ then lie in the extension $K = \mathbb{Q}(\cos 2\pi/N)$, as follows from the characterization (2) (with n in that formula equal to N here). More precisely, if k is odd, $\cos 2\pi k/N$ is a conjugate of $\cos 2\pi/N$ in K. If k is even, $\cos 2\pi k/N$ equals $-\cos 2\pi k'/N$, for $k' = N/2-k$ odd; therefore it is a conjugate of $-\cos 2\pi/N$. Either way, $\cos 2\pi k/N$ lies in K, and therefore so does $s(k)$, since f is a rational function.

The operation of taking conjugates commutes with applying f (monomorphisms preserve sums, products and inverses, and fix the numbers 1 and 2). Putting this together with the previous paragraph, we conclude that half of the $s(k)$ (those where k is odd) make up the conjugates in K of $s(1)$, while the other half make up the conjugates of $s(2)$ (taking $k = 2$ as a representative of the even k's). It follows that

$$\sum_{k=1}^{N/2-1} s(k) = \mathrm{Tr}_K s(1) + \mathrm{Tr}_K s(2) = \mathrm{Tr}_K f\left(\cos\frac{2\pi k}{N}\right) + \mathrm{Tr}_K f\left(\cos\frac{2\times 2\pi k}{N}\right).$$

Thus $S(n, m)$ is the sum of two traces of algebraic numbers, and so rational.

Now let $N/2$ be arbitrary. Our strategy is the same: we partition the values of k according to their gcd with N. Let $d_1, \ldots, d_l$ be all the divisors of N apart from N and $N/2$, and define

$$D_i := \{k : \gcd(k, N) = d_i,\ 0 < k < N/2\} = \{d_i j : \gcd(j, N/d_i) = 1,\ 0 < j < N/(2d_i)\}.$$

The D_i are disjoint, and together they account for all the k in the sum (3). Moreover,

$$\sum_{k\in D_i} s(k) = \sum_{\substack{j:\gcd(j,N/d_i)=1\\ 0<j<N/(2d_i)}} f\left(\cos\frac{2\pi j}{N/d_i}\right) = \mathrm{Tr}_{\mathbb{Q}(\cos\frac{2\pi}{N/d_i})} f\left(\cos\frac{2\pi}{N/d_i}\right),$$

where the last equality follows from the same reasoning used earlier for k odd (with N replaced by N/d_i). We have expressed $S(n, m)$ as a sum of traces of algebraic numbers, which means it is rational. □

This result allows us to prove that the generating function for the sequence $\{S(n,m)\}_{m>0}$ (for fixed odd $n>0$) is a rational function.

Lemma 4. *If $n>1$ is odd, $m\geq 1$, and*

$$F_n(x)=\sum_{m=0}^{\infty} S(n,m)x^m,$$

then there exist $P(x)$, $Q(x)\in\mathbb{Z}[x]$ such that $F_n(x)=P(x)/Q(x)$.

Proof. Using the formula for the sum of a geometric series, we write

$$F_n(x)=\sum_{m=0}^{\infty}\left(\sum_{k=1}^{(n-1)/2}\sec^{2m}\frac{k\pi}{n+1}\right)x^m=\sum_{k=1}^{(n-1)/2}\frac{1}{1-x\sec^2\frac{k\pi}{n+1}},$$

so that

$$Q(x)=\prod_{k=1}^{(n-1)/2}\left(1-x\sec^2\frac{k\pi}{n+1}\right).$$

We will show that $Q(x)$ is a polynomial with rational coefficients. Set

$$b_k:=\sec^2\frac{k\pi}{n+1},$$

where $1\leq k\leq (n-1)/2$. Let s_i be the sum of the products of each i-element subset of the set $\{b_1,b_2,\ldots,b_{(n-1)/2}\}$ (in other words, s_i is the i-th elementary symmetric polynomial applied to the b_i). The coefficient of x^i in $Q(x)$ is $(-1)^i s_i$. Also, let

$$p_r:=\sum_{k=1}^{n} b_k^r.$$

The Newton–Girard formulas tell us that

$$p_i-s_1p_{i-1}+s_2p_{i-2}+\cdots+(-1)^{i-1}s_{i-1}p_1+(-1)^i i s_i=0,$$

for all $1\leq i\leq (n-1)/2$. It is clear that p_i is rational for all i by Lemma 4. An easy induction argument implies that s_i is rational for all i. Since the coefficients of $Q(x)$ can be expressed in terms of the s_i, we see that $Q(x)$ has rational coefficients. Thus $P(x)=F_n(x)Q(x)$ has rational coefficients. The desired result follows. □

Lemma 5. *Suppose that a and b are algebraic numbers, and*

$$F(x)=\frac{a}{1-bx}=\sum_{n=0}^{\infty}a_nx^n.$$

If a_n is an algebraic integer for all n, then b is an algebraic integer.

Proof. The assumption implies that $a_n = ab^n$. We know that ab^n is an algebraic integer for all n, and so lies in the ring of algebraic integers of the field $K = \mathbb{Q}(b)$. This ring is finitely generated as an abelian group. Suppose that it is generated by $\{v_1, v_2, \ldots, v_l\}$. Then b^n must be in the finitely generated abelian group generated by $\{v_1/a, \ldots, v_l/a\}$ for all n. Lemma 2.8 of [ST] states that a complex number θ is an algebraic integer if and only if the additive group generated by all powers $1, \theta, \theta^2, \ldots$ is finitely generated. Thus, b is an algebraic integer. □

Now, we wish to expand upon the ideas presented in Lemma 5.

Definition. A sequence $\{a_n\}$ of algebraic numbers has a *bounded denominator* if there exists a positive integer m such that ma_n is an algebraic integer for all n.

Lemma 6. *Let*

$$F(x) = \sum_{n=0}^{\infty} a_n x^n,$$

where $\{a_n\}$ is a sequence with bounded denominator. Suppose $p(x)$ is a polynomial whose coefficients are algebraic numbers and let

$$F(x)\,p(x) = \sum_{n=0}^{\infty} b_n x^n.$$

Then, the sequence $\{b_n\}$ has bounded denominator.

Proof. This follows from the fact that the algebraic numbers form a subfield of the complex numbers and the fact that given an algebraic number a there exists a positive integer n such that na is an algebraic integer. □

Lemma 7. *Let $\zeta_{4p} = e^{2\pi i/4p}$, where p is an odd prime. Then*

$$N_{\mathbb{Q}(\zeta_{4p})}(\zeta_{4p} + \zeta_{4p}^{-1}) = p^2.$$

Proof. First note that

$$\zeta_{4p} + \zeta_{4p}^{-1} = 2\cos\frac{\pi}{2p},$$

and recall the characterization of the conjugates of $\cos 2\pi/n$ given in (2). We have

$$N_{\mathbb{Q}(\zeta_{4p})}(\zeta_{4p} + \zeta_{4p}^{-1}) = \prod_{\substack{(k,4p)=1\\ 1\le k\le 4p}} \left(e^{\frac{2\pi ik}{4p}} + e^{\frac{-2\pi ik}{4p}}\right) = \zeta_{4p}^{-\phi(4p)2p} \prod_{\substack{(k,4p)=1\\ 1\le k\le 4p}} \left(e^{\frac{4\pi ik}{4p}} + 1\right).$$

Now, we know that

$$N_{\mathbb{Q}(\zeta_{2p})}(\zeta_{2p} + 1) = \prod_{\substack{(k,2p)=1\\ 1\le k\le 2p}} (e^{\frac{2\pi ik}{2p}} + 1).$$

This implies

$$\zeta_{4p}^{-\phi(4p)2p} \prod_{\substack{(k,4p)=1 \\ 1\le k\le 4p}} \left(e^{\frac{4\pi ik}{4p}} + 1\right) = N_{\mathbb{Q}(\zeta_{2p})}(\zeta_{2p}+1)^2(e^{-2\pi i}).$$

Now, the minimal polynomial of ζ_{2p} is the same as that of $-\zeta_p$. Furthermore,

$$r(x) = x^{p-1} + x^{p-2} + \cdots + x + 1 = \prod_{k=1}^{p-1}(x - \zeta_p^k).$$

So, $N_{\mathbb{Q}(\zeta_p)}(1-\zeta_p) = r(1) = p$ since the minimal polynomial of ζ_p is $r(x)$. Thus,

$$N_{\mathbb{Q}(\zeta_{2p})}(\zeta_{2p}+1)^2(e^{-2\pi i}) = N_{\mathbb{Q}(\zeta_p)}(1-\zeta_p)^2 = p^2,$$

as desired. □

Lemma 8. *If, for all k satisfying $1 \le k \le (n-1)/2$, the value of $\sec^2 \dfrac{k\pi}{n+1}$ is an algebraic integer, then $n+1$ is a power of two.*

Proof. Assume that $n+1$ is not a power of two. Let p be an odd prime factor of $2(n+1)$. Since n is odd, $2(n+1)$ is a multiple of 4 and so $4p$ divides $2(n+1)$. Let $k = 2(n+1)/(4p)$, so $2(n+1)/k = 4p$. Then

$$\sec^2 \frac{k\pi}{n+1} = \left(\frac{2}{\zeta_{2(n+1)}^k + \zeta_{2(n+1)}^{-k}}\right)^2 = \left(\frac{2}{\zeta_{4p} + \zeta_{4p}^{-1}}\right)^2.$$

Now, from the previous lemma, $N_{\mathbb{Q}(\zeta_{4p})}(\zeta_{4p} + \zeta_{4p}^{-1}) = p^2$. This implies

$$N_{\mathbb{Q}(\zeta_{4p})}\left(\frac{2}{\zeta_{4p} + \zeta_{4p}^{-1}}\right)^2 = \frac{2^{2\phi(4p)}}{p^4}.$$

Then, since p is an odd prime we know that $2^{2\phi(4p)}/p^4$ is not an integer. This means that with the chosen k, $\sec^2 k\pi/(n+1)$ is not an algebraic integer. This proves the desired result. □

4. Proof of the theorems

Proof of Theorem 2. This is a more general version of Lemma 5. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be all the numbers whose reciprocals are zeros of $Q(x)$. Then $F(x)$ has a partial fraction expansion whose terms are of the form

$$\frac{A_{i,l}}{(1-\alpha_i x)^l},$$

plus a polynomial part. Write

$$Q(x) = \prod_{i=1}^{n} (1 - \alpha_i x)^{k_i}.$$

Let j be the largest positive integer such that in the partial fraction decomposition of $F(x)$ the term $A_{i,j}/(1-\alpha_i x)^j$ is nonzero. Clearly $j > 0$, since $P(x)$ and $Q(x)$ are relatively prime. Now, let

$$Q_i(x) = \frac{Q(x)}{(1-\alpha_i x)^{k_i - j + 1}}.$$

The highest power of $(1-\alpha_i x)$ that divides $Q_i(x)$ is clearly $j-1$.

We have

$$F(x)Q_i(x) = \sum_{n=0}^{\infty} b_n x^n.$$

Then, by Lemma 6, $\{b_n\}$ has a bounded denominator. Now, we will consider the effect of multiplying $F(x)$ and $Q_i(x)$ by considering what happens to each term in the partial fraction expansion of $F(x)$. With the exception of the term

$$\frac{A_{i,j}}{(1-\alpha_i x)^j},$$

$Q_i(x)$ times a term in the partial fraction expansion of $F(x)$ is a polynomial of finite degree. Now, one can see that

$$Q_i(x)\frac{A_{i,j}}{(1-\alpha_i x)^j} = \frac{Q_i(x)}{(1-\alpha_i x)^{j-1}}\frac{A_{i,j}}{(1-\alpha_i x)}.$$

It is clear that $Q_i(x)/(1-\alpha_i x)^{j-1}$ is a polynomial. Thus,

$$F(x)Q_i(x) = q(x) + \frac{D_i}{1-\alpha_i x},$$

where $q(x)$ is a polynomial and D_i is some algebraic number. So, we can say that for sufficiently large n, $b_n = D_i \alpha_i^n$ where D_i and b_n are algebraic numbers. Then, by Lemma 5, α_i is an algebraic integer. □

Proof of Theorem 1. Suppose $S(n,m)$ is an integer for all $m > 0$. By Lemma 4,

$$F_n(x) = \sum_{m=0}^{\infty}\left(\sum_{k=1}^{(n-1)/2} \sec^{2m}\frac{k\pi}{n+1}\right)x^m = \sum_{k=1}^{(n-1)/2}\left(\frac{1}{1 - x\sec(\frac{k\pi}{n+1})}\right)$$

is a rational function. Hence, $F_n(x) = P(x)/Q(x)$ where $P(x)$, $Q(x) \in \mathbb{Q}[x]$. Theorem 1 now implies that $\sec^2(k\pi/(n+1))$ is an algebraic integer for all k with $1 \le k \le (n-1)/2$. According to Lemma 8, this means $n+1$ is a power of two. □

Acknowledgements

The author thanks his advisor, Jeremy Rouse, for introducing him to this problem and keeping him motivated throughout the project. The author also thanks the anonymous referee for helpful comments.

References

[AMM 1983] M. Larsen, "Problems and solutions: E 2993", *Amer. Math. Monthly* **90**:4 (1983), 287.

[AMM 1986] M. Larsen, "Solution to problem E 2993: An application of Newton's formulae", *Amer. Math. Monthly* **93**:6 (1986), 483.

[AMM 2006] AMM, "Problems and solutions", *Amer. Math. Monthly* **113** (2006), 268.

[AMM 2008] S. Rabinowitz and NSA Problems Group, "Problems and solutions. Solutions: sometimes an integer: 11213(a)", *Amer. Math. Monthly* **115**:4 (2008), 366–367.

[Milne 2009] J. S. Milne, *Algebraic number theory* (version 3.00), 2009.

[Stewart and Tall 2002] I. Stewart and D. Tall, *Algebraic number theory and Fermat's last theorem*, 3rd ed., A K Peters, Natick, MA, 2002. MR 2002k:11001 Zbl 0994.11001

Received: 2010-07-19 Revised: 2011-02-01 Accepted: 2011-02-02

mudrock2@illinois.edu *Mathematics Department,*
University of Illinois at Urbana-Champaign,
1409 West Green Street, Urbana, IL 61801, United States

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.127

Clique-relaxed graph coloring

Charles Lundon, Jennifer Firkins Nordstrom, Cassandra Naymie, Erin Pitney, William Sehorn and Charlie Suer

(Communicated by Vadim Ponomarenko)

We define a generalization of the chromatic number of a graph G called the k-clique-relaxed chromatic number, denoted $\chi^{(k)}(G)$. We prove bounds on $\chi^{(k)}(G)$ for all graphs G, including corollaries for outerplanar and planar graphs. We also define the k-clique-relaxed game chromatic number, $\chi_g^{(k)}(G)$, of a graph G. We prove $\chi_g^{(2)}(G) \leq 4$ for all outerplanar graphs G, and give an example of an outerplanar graph H with $\chi_g^{(2)}(H) \geq 3$. Finally, we prove that if H is a member of a particular subclass of outerplanar graphs, then $\chi_g^{(2)}(H) \leq 3$.

1. Introduction

The *chromatic number* of a graph G, denoted $\chi(G)$, is the least number of colors required to color the vertices of G such that adjacent vertices receive different colors. The study of this characteristic of graphs is interesting in itself, and several extensions have also been explored. For example, the *k-relaxed chromatic number* of a graph G, denoted $\chi^k(G)$, is the least number of colors necessary to color the vertices of G such that each vertex is adjacent to at most k vertices of the same color. Note that $\chi^0(G) = \chi(G)$. This parameter has been studied in many papers, including [Cowen et al. 1986; 1997; Eaton and Hull 1999]. In this paper we introduce a relaxation to vertex coloring which forbids monochromatic $(k+1)$-cliques, where a *k-clique* is a set of k pairwise-adjacent vertices.

Another area of research branching from graph coloring is competitive graph coloring. Two players, Alice and Bob, take turns (with Alice going first) coloring uncolored vertices of a graph G with legal colors from a set X of m colors, where the definition of a legal color for a vertex varies depending on the version of the game. In the standard game [Bodlaender 1992], a color $\alpha \in X$ is *legal* for an uncolored vertex u if u has no neighbors already colored α. Alice wins this game if all vertices of G are eventually colored. Bob wins when there is an uncolored

MSC2000: 05C15.
Keywords: competitive coloring, outerplanar graph, clique, relaxed coloring.
Partially supported by NSF grant DMS 0649068.

vertex for which no legal color exists. The least m such that Alice has a winning strategy for this game is called the *game chromatic number of* G, and is denoted $\chi_g(G)$. In the k-relaxed version of the game [Chou et al. 2003; Dunn and Kierstead 2004a; 2004b; 2004c; He et al. 2004], a color is legal for a vertex if it does not result in any vertex with more than k neighbors of the same color. Said differently, a color $\alpha \in X$ is legal for an uncolored vertex u if once u is colored α, for every $\beta \in X$, the subgraph H induced by all vertices colored β satisfies $\Delta(H) \leq k$, where $\Delta(H)$ is the maximum degree of H. Alice wins if all the vertices of G are eventually colored. Bob wins if there is at least one uncolored vertex in G with no legal color. The least m such that Alice has a winning strategy for this game is called the *k-relaxed game chromatic number of* G, denoted $\chi_g^k(G)$. We will show how competitive coloring can be integrated with the definition of a clique-relaxed coloring.

2. Clique-relaxed coloring

A coloring of a graph G is a proper k-clique-relaxed coloring if G has no monochromatic $(k+1)$-cliques. For any graph G, the *k-clique-relaxed chromatic number of* G, denoted $\chi^{(k)}(G)$, is defined as the least number of colors that can be used to color the vertices of a graph G such that if H is a subgraph induced by one of the color classes, then $\omega(H) \leq k$, where $\omega(H)$ is the size of the largest clique in H. Notice that $\chi^{(1)}(G) = \chi(G)$ for all graphs G, and more generally that for every positive integer k, we have that $\chi^{(k)}(G) \leq \chi^{k-1}(G)$. The following theorem gives an upper bound for the k-clique-relaxed chromatic number of a graph G in terms of the standard chromatic number of G.

Theorem 1. *Let G be a graph. Then $\chi^{(k)}(G) \leq \left\lceil \dfrac{\chi(G)}{k} \right\rceil$ for any positive integer k.*

Proof. Let G be a graph with $\chi(G) = m$. Then G has a proper m-coloring. Let k be a positive integer. We know that there are unique nonnegative integers q and r, $r < k$, such that $m = qk + r$. We can thus divide the m colors into q groups of size k and one of size r if $r \neq 0$. This gives $\lceil m/k \rceil = n$ groups. Let $A_1, A_2, \ldots, A_n$ be these groups of colors. Now, using the proper m-coloring, we color a vertex v with a color β_i if $c(v) \in A_i$, where $c(v)$ denotes the color of v. The colors used in this new coloring are $\beta_1, \beta_2, \ldots, \beta_n$. Thus n colors are used. Notice that the vertices of any $(k+1)$-clique in the proper m-coloring must have been colored using $k+1$ different colors, and any set of $k+1$ colors from the proper m-coloring must be in at least two groups A_i and A_j where $i \neq j$. So in the new coloring, the vertices of any $(k+1)$-clique must include at least two colors. Therefore, the new coloring is a proper k-clique-relaxed coloring with n colors. So $\chi^{(k)}(G) \leq n = \lceil \chi(G)/k \rceil$. □

Using known characteristics of outerplanar and planar graphs it is easy to apply the result in Theorem 1 to these classes of graphs. By the 2-degeneracy of

outerplanar graphs, $\chi(G) \leq 3$ for all outerplanar graphs G, and by the four-color theorem [Appel and Haken 1976], $\chi(H) \leq 4$ for all planar graphs H. We then have the following corollary.

Corollary 2. *If G is an outerplanar graph, then*

$$\chi^{(2)}(G) \leq 2 \qquad \textit{and} \qquad \chi^{(k)}(G) = 1 \quad \textit{for } k \geq 3.$$

Similarly, if H is a planar graph, then

$$\chi^{(2)}(H) \leq 2, \qquad \chi^{(3)}(H) \leq 2, \qquad \textit{and} \qquad \chi^{(k)}(H) = 1 \quad \textit{for } k \geq 4.$$

Observe that K_3 is an outerplanar graph with $\chi^{(2)}(K_3) = 2$, since if every vertex in K_3 is colored α, there is a 3-clique in the subgraph induced by the color α. Similarly, K_4 is a planar graph with $\chi^{(2)}(K_4) = 2$ and $\chi^{(3)}(K_4) = 2$, Since if every vertex in K_4 is colored α, there is a 4-clique and four 3-cliques in the subgraph induced by the color α.

We note that our discussion of clique-relaxed coloring can be reframed within the context of hypergraph colorings. For a given graph G we define the hypergraph $H = (V, E)$, where $V = V(G)$ and E is the set of hyperedges induced by the $(k+1)$-cliques in G. In this way, k-clique-relaxed coloring in G is equivalent to standard hypergraph coloring in H. However, for the simplicity of our arguments, we will remain within the context of graphs rather than hypergraphs.

3. Clique-relaxed coloring game

A natural extension of this relaxed coloring number is its application to competitive graph coloring. The *k-clique-relaxed n-coloring game* on a graph G is between two players, Alice and Bob, who take turns coloring uncolored vertices of G with colors from a set X of n colors. A color $\alpha \in X$ is *legal* for an uncolored vertex u if coloring u with α does not result in a monochromatic $(k+1)$-clique. At each step the players must color an uncolored vertex with a legal color. As before with the k-relaxed coloring game, we can restate this in terms of the subgraphs induced by the color classes. A color α is legal for u if once u is colored α, for every $\beta \in X$, the subgraph H induced by the vertices of color β satisfies $\omega(H) \leq k$. Alice always colors first, and she wins the game when all the vertices are colored. Hence, Bob wins when there is at least one uncolored vertex in G with no legal color. The *k-clique-relaxed game chromatic number of G*, denoted $\chi_g^{(k)}(G)$, is the least n such that Alice has a winning strategy in the k-clique-relaxed n-coloring game on G.

Notice that $\chi_g^{(1)}(G) = \chi_g(G)$ for all graphs. Also, since outerplanar graphs have maximum clique size at most three, $\chi_g^{(k)}(G) = 1$ for all outerplanar graphs G and $k \geq 3$. Therefore, we will be concerned only with the 2-clique-relaxed game on

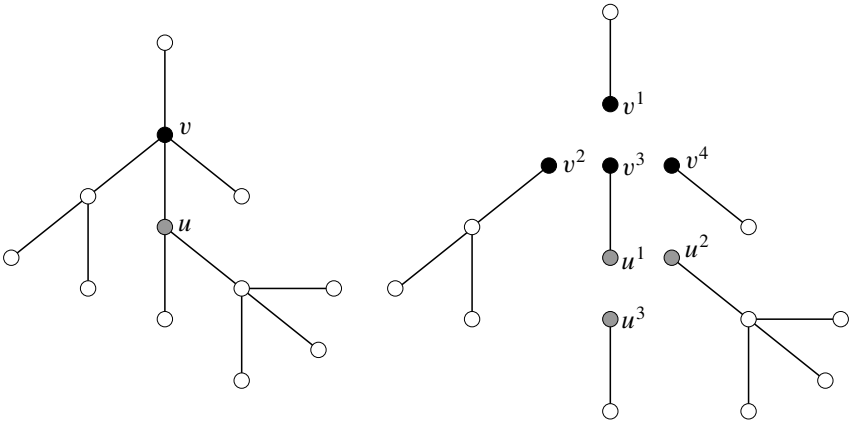


Figure 1. Alice will create trunks at the vertices v and u.

outerplanar graphs. Before proving an upper bound for the 2-clique-relaxed game chromatic number of outerplanar graphs, we will reprove Lemma 1 of [Guan and Zhu 1999] which is key to Alice's strategy.

The *separator strategy* on a tree T is defined as follows. Let $c(v)$ denote the color of a vertex v. At any point in the coloring game on T if a vertex v is colored and has degree d, Alice will imagine vertex v is replaced by d vertices, all colored $c(v)$, where each of these d vertices is incident with exactly one edge that was incident with v. We call these partially-colored subgraphs *trunks*. For example, consider the partially-colored tree on the left side of Figure 1. The vertices v and u are colored, so Alice creates trunks at these vertices as shown on the right side of the figure.

Lemma 3. *Using the separator strategy, Alice can ensure that after each of her turns each trunk has at most two colored vertices.*

Proof. Let T be a tree. It is clear that the property holds after Alice's first turn. Suppose this holds after Alice's k-th turn, and Bob colors a vertex u on a trunk. So at the end of Bob's turn there is at most one trunk with more than two colored vertices. If such a trunk exists, it is the trunk with vertex u, and this trunk has three colored vertices. If u lies on the path between the other two colored vertices, then according to Alice's view of the game, this trunk will be broken into two trunks, each with two colored vertices. Then, if possible, Alice will color on a trunk with only one colored vertex. If there are no such trunks, she can color a vertex on the distinct path between two colored vertices within a trunk with two colored vertices, separating the trunk at the vertex she just colored. If u does not lie on this path, Alice can color the unique vertex at which the paths between the three colored vertices intersect. Call this vertex v. As she is using the separator

strategy, she then separates the unique trunk containing v of T at v into d trunks where $d = \deg(v)$. Now each of these d trunks has at most two colored vertices.

Suppose, instead, that Bob colors in a trunk with only one colored vertex. Then Alice plays as above in the case when Bob colored on the path between two colored vertices. Thus, in either case, the property holds after Alice's $(k+1)$-th turn. □

Theorem 4. *Let G be an outerplanar graph. Then $\chi_g^{(2)}(G) \leq 4$.*

Proof. Let $G = (V, E)$ be an outerplanar graph. Alice will use a strategy for the 2-clique-relaxed 4-coloring game on G adapted from [Guan and Zhu 1999]. Alice begins by creating auxiliary graphs G' and T which she will use to determine which vertex she colors in the game on G.

To create $G' = (V', E')$, Alice adds edges to G so that G' is maximally outerplanar. Notice that $V = V'$, and $E \subseteq E'$. Guan and Zhu [1999] showed that for every maximally outerplanar graph, there is a linear ordering $L = v_1 v_2 \dots v_n$ of the vertices of G' such that

- v_1 and v_2 are adjacent,
- $v_1 v_2$ is on the outer face of G', and
- for all $i \geq 3$, v_i is adjacent to exactly two vertices $v_{a(i)}$ and $v_{b(i)}$ such that $a(i) < i$ and $b(i) < i$.

We call $v_{a(i)}$ and $v_{b(i)}$ the *major parent* and *minor parent* of v_i, respectively, where $a(i) < b(i)$.

To create $T = (V_T, E_T)$, Alice deletes all edges of the form $v_i v_{b(i)}$. In other words, for each vertex u she deletes the edge between u and its minor parent. According to Lemma 1 of [Guan and Zhu 1999], each vertex is the minor parent of at most two vertices. Since each vertex also has at most one minor parent, every vertex in T is incident to at most three deleted edges from G'. Notice that $V_T = V' = V$ and $E_T \subseteq E'$.

We can see in T that v_1 and v_2 are still adjacent, and now for all $i \geq 3$, v_i is adjacent to exactly one vertex with a lower index, namely its major parent $v_{a(i)}$. So T is a tree. Alice will use the separator strategy on T to choose which vertex she will color. Let v be the vertex she chooses. She will look at the partially colored graph G and choose a legal color for v. We show that in the 2-clique-relaxed 4-coloring game, v will always have a legal color.

We proved in Lemma 3 that by using the separator strategy, Alice can ensure that after her turn each trunk has at most two colored vertices. After Bob's turn there may be one trunk with three colored vertices, so v is adjacent to at most three colored vertices in T. Since, as noted earlier, each vertex is incident to at most three deleted edges from G', the vertex v may be adjacent to three additional colored vertices in G'. Since $E_T \subseteq E'$, v is adjacent to at most six colored vertices

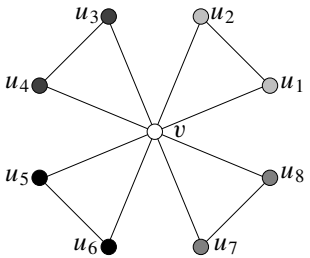


Figure 2. The vertex v is uncolorable in the 2-clique-relaxed 4-coloring game.

in G'. Also, because $E \subseteq E'$, we know that v is adjacent to at most six colored vertices in G. If v is uncolorable, then it must form a 3-clique with each of the four color classes (see Figure 2). Thus, it must be adjacent to at least eight colored vertices in G. Since v is only adjacent to six colored vertices, there is a legal color for v and Alice can win the 2-clique-relaxed 4-coloring-game on G. □

We do not yet know if the above bound is sharp. The theorem that follows gives an example of a graph G such that $\chi_g^{(2)}(G) \geq 3$. In order to prove this we show that Bob has a winning strategy in the 2-clique-relaxed 2-coloring game on G. This means that Alice would need three or more colors to have a winning strategy on G. We begin our proof with two lemmas which involve subgraphs of G.

Lemma 5. *Let H be the partially colored graph in Figure 3, where $c(v_1) = c(v_2)$, $c(v_3) \neq c(v_1)$, the vertices x, y, z, and w are uncolored, the color $c(v_1)$ is legal for both x and z, and the color $c(v_2)$ is legal for both y and w. If H is a subgraph of an outerplanar graph G at any point in the 2-clique-relaxed 2-coloring game on G, then Bob has a winning strategy.*

Proof. Assume v_1 and v_2 are colored α and v_3 is colored β. If it is Bob's turn he can color either y or w with β. Vertex z can then be colored neither α nor β, so Bob wins. Suppose instead that it is Alice's turn. If she does not color z, then either y or w is still uncolored after her turn (if not both). Suppose without loss of

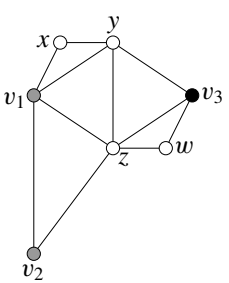


Figure 3. Bob can win the 2-clique-relaxed 2-coloring game.

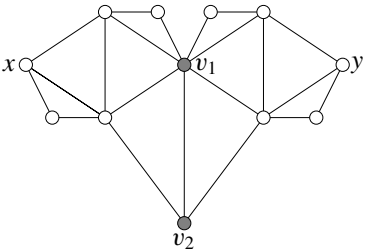


Figure 4. Bob can win the 2-clique-relaxed 2-coloring game.

generality that y is uncolored. Then Bob can color y with β leaving z uncolorable. If Alice does color z, she must color it β. Bob can then color x with α. Now y can be colored neither α nor β, so Bob has a winning strategy. □

Lemma 6. *Let H be the partially colored graph in Figure 4, where $c(v_1) = c(v_2) = \alpha$ and all other vertices in the subgraph are uncolored. Suppose Alice and Bob are playing the 2-clique-relaxed 2-coloring game on an outerplanar graph G with colors α and β. If H is a subgraph of G and β is legal for both $x$and y, then Bob has a winning strategy.*

Proof. Assume v_1 and v_2 are colored α. If it is Bob's turn he can color either x or y with β, and by Lemma 5 Bob can win. If instead it is Alice's turn, she can only play on one side of the line of symmetry. If she colors a vertex on the side with x, Bob can color y with β; if she colors a vertex on the side with y, Bob can color x with β. Either way, by Lemma 5, Bob can win. □

Theorem 7. *There exists an outerplanar graph G such that $\chi_g^{(2)}(G) \geq 3$.*

Proof. Consider the graph in Figure 5. If Alice colors v with α, then Bob can color u_1, u_2, or u_3 with α, and, by Lemma 6, Bob can win. If Alice does not color v,

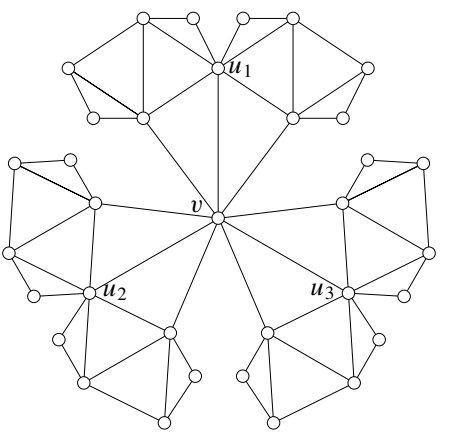


Figure 5. Bob has a winning strategy on this graph.

then Bob can color v on his first turn with α. On Bob's second turn at least one of the three identical trunks adjacent to v has no colored vertices since Alice has only played twice. Suppose without loss of generality that the part containing u_1 has no colored vertices. Bob can color u_1 with α, and by Lemma 6 he can win. □

4. Family representation for outerplanar graphs

In this section, we present a representation for outerplanar graphs such that each component of the graph is rooted, and its vertices are organized into generations. Recall that a graph is outerplanar if and only if it has no $K_{2,3}$ or K_4 minor. Let G be an outerplanar graph with m components, and let $G_1, G_2, \ldots, G_m$ be the components of G.

- For each G_i choose any vertex r_i to be the root.
- Partition $V(G_i)$ into $V_0^i, V_1^i, \ldots, V_k^i$ such that

$$V_j^i = \{x \in V(G) \mid d(x, r_i) = j\},$$

where $d(x, r_i)$ is the distance between x and r_i. Define $V_j = \bigcup_{i=1}^{m} V_j^i$. Each V_j is the *j-th generation* of G.

Since the vertex set of any outerplanar graph can be partitioned according to the distance of a vertex from a fixed root and the edge set remains unchanged, all outerplanar graphs have a family representation.

Let $v \in V_j$ for some $j \geq 1$. Then u is a *parent* of v if $u \in V_{j-1} \cap N(v)$, where $N(v)$ is the set of neighbors of v. Likewise, u is a *child* of v if $u \in V_{j+1} \cap N(v)$. We call a vertex u a *descendant* of v if there is a shortest (nonempty) path from u to the root that includes v. We note that if the following properties of the family representation are true for each component of G, then they are true for G; thus, we may assume that G is connected.

Proposition 8. *All vertices in G have at most two parents.*

Proof. Assume that a vertex $x \in V_j$ has three parents in V_{j-1}. Note, $j \neq 1$ since V_0 has only one vertex. Call the three parents a, b, and c. Let $M = \{V_i \mid i < j - 1\}$. Clearly, $G[M]$, the graph induced by M, is connected. The vertices a, b, and c each have at least one parent in M. Let $X = \{a, b, c\}$ and let $Y = \{x, G[M]\}$. These bipartite sets and the edges that connect them form a minor of $K_{2,3}$, contradicting the fact that G is outerplanar. Thus, each vertex has at most two parents. □

Proposition 9. *For each* $v \in V_j$, $|N(v) \cap V_j| \leq 2$.

Proof. Assume that a vertex $x \in V_j$ has three neighbors in V_j. Call the three neighbors a, b, and c. Let $M = \{V_i \mid i < j\}$. As in the previous proof, we see that with $X = \{a, b, c\}$ and $Y = \{x, G[M]\}$, we have a $K_{2,3}$ minor, contradicting the

fact that G is outerplanar. So, each vertex in the j-th generation has at most two neighbors in the j-th generation. □

5. The coloring game on certain outerplanar graphs

It is known [Guan and Zhu 1999] that for the class $\mathcal{G}$ of outerplanar graphs, that

$$6 \leq \max_{G \in \mathcal{G}} \chi_g(G) \leq 7.$$

In this section we consider a specific subclass of outerplanar graphs for which we can improve this upper bound. We consider outerplanar graphs for which there exists a family representation such that each vertex u has at most one parent $p(u)$. This means that for each vertex $v \in V(G_i)$ there is a unique shortest path from v to root r_i. We call this class $\mathcal{F}$. See Figure 6.

Alice will use an *activation strategy* to win the usual 6-coloring game and the 2-clique-relaxed 3-coloring game on graphs in $\mathcal{F}$. At any point in the game, we define U to be the set of uncolored vertices, and C to be the set of colored vertices. Alice maintains a set of active vertices, A. Any colored vertex is automatically active, and once a vertex is active it remains active. Therefore $C \subseteq A$.

Activation strategy: On Alice's first turn, she colors a vertex in V_0. Suppose Bob colors vertex $v \in V(G_i)$.

(1) Search stage:

- If v is not a root and $p(v)$ is uncolored, Alice begins activating vertices along the shortest path from v to root r_i. As she does this, there are four possible cases for each vertex x she reaches.
  - If x is active and uncolored, she lets $u = x$ and moves to the coloring stage.
  - If x is inactive and is the root r_i, she activates x, chooses $u = x$, and moves to the coloring stage.

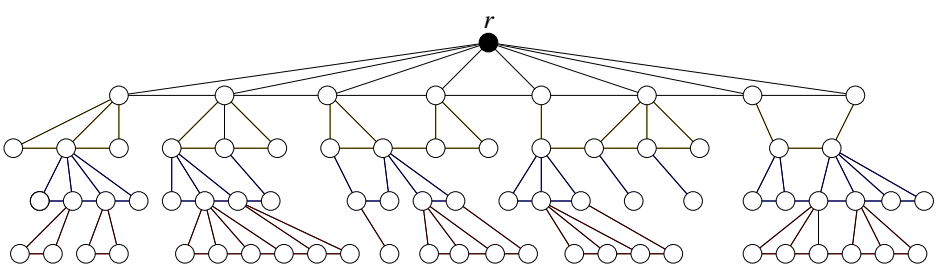


Figure 6. An example of a graph in $\mathcal{F}$.

  - If x is inactive and $p(x)$ is colored, she activates x, chooses $u = x$, and moves to the coloring stage.
  - If x is inactive and $p(x)$ is uncolored, she activates x and continues up the path.
- If v is a root or $p(v)$ is colored, Alice chooses an arbitrary uncolored vertex $u \in V_j$, where j is the least index such that V_j has an uncolored vertex, and moves to the coloring stage.

(2) Coloring stage:

- On each turn, Alice chooses a legal color for u.

We now prove an important lemma which will help bound the parameters of interest.

Lemma 10. *If Alice uses the activation strategy, at any point in the game any uncolored vertex u has at most two active children.*

Proof. Consider the case where u has no active children. The strategy ensures that Alice will not color a descendant of any inactive vertex. Thus, if Alice activates a child of u, it must be the direct result of Bob coloring a descendant of u. When Alice activates a child of u, she activates u as well. Now consider Alice activating a second child of u. Again, this must be a result of Bob coloring a descendant of u by the argument above. After Alice activates the second child of u, she will take action at u. Since u is active, Alice colors u. Therefore, an uncolored vertex u has at most two active children. □

Theorem 11. *For all graphs G in $\mathcal{F}$, $\chi_g(G) \leq 6$.*

Proof. Consider an uncolored vertex u. Note that u has at most one parent $p(u)$, and by Proposition 9, u has at most two adjacent siblings, say u^* and u'. See Figure 7. It is easy to see that if Alice uses the activation strategy, it may be the case that $p(u)$, u^*, and u' are all colored with u remaining uncolored. Since u has at most two active children, it has at most two colored children. Therefore, u has at most five colored neighbors. This means that Alice needs at most six colors available to win the original game on graphs in $\mathcal{F}$. □

Now we prove a similar result for the 2-clique-relaxed game. Recall that in Theorem 4 and Theorem 7 we showed that

$$3 \leq \max_{G \in \mathcal{G}} \chi_g^{(2)}(G) \leq 4$$

for the class of outerplanar graphs $\mathcal{G}$. With the following result, we provide an improved upper bound for the class $\mathcal{F}$.

Corollary 12. *For all graphs G in $\mathcal{F}$, $\chi_g^{(2)}(G) \leq 3$.*

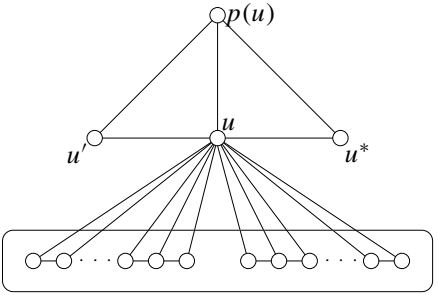


Figure 7. An uncolored vertex u and its neighbors.

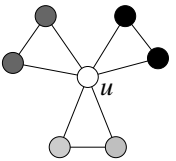


Figure 8. The vertex u is uncolorable in the 2-clique-relaxed 3-coloring game.

Proof. Suppose Alice and Bob are playing the 2-clique-relaxed coloring game on a graph in $\mathcal{F}$ with three colors. For Bob to win the game, he requires an uncolored vertex u, with neighbors as in Figure 8. This would require that six of the neighbors of u be colored while u remains uncolored. By the proof for Theorem 11, u has at most five colored neighbors. Hence, three colors are sufficient for Alice to win the 2-clique relaxed game on any graph in $\mathcal{F}$. □

6. Future work

At present, we do not know whether the bounds in Theorem 11 and Corollary 12 are tight. In the case of the latter, it is clear that the graph in Theorem 7 is not in $\mathcal{F}$. Showing this bound is tight would require providing an example of a graph in $\mathcal{F}$ such that Bob has a winning strategy with 2 colors. However, it may be the case that Alice has a winning strategy with 2 colors. We are certain that the strategy we have provided will not suffice; however, it is possible that a modification could yield an upper bound of 2.

We now have an upper bound, $\chi_g^{(2)}(G) \leq 4$, for outerplanar graphs G and an example of an outerplanar graph such that $\chi_g^{(2)}(G) \geq 3$. The next question is whether there exists an outerplanar graph G such that $\chi_g^{(2)}(G) = 4$. If there is, then such an example must lie outside of the subclass $\mathcal{F}$ of outerplanar graphs. In particular, Proposition 8 guarantees that such an example would require a vertex with two parents.

Another area for further investigation is the clique-relaxed game chromatic number of planar graphs. All planar graphs have maximum clique size at most four. For this reason, with a k-clique relaxation, where $k \geq 4$, planar graphs can always be completely colored with one color. The games of interest are then the 2- and 3-clique-relaxed games on planar graphs.

More broadly, as we noted earlier in Section 2, much of this work can be reframed in terms of hypergraph coloring. We have presented competitive coloring

results for a specific class of hypergraphs. This could lead to more questions in the area of competitive hypergraph coloring.

References

[Appel and Haken 1976] K. Appel and W. Haken, "Every planar map is four colorable", *Bull. Amer. Math. Soc.* **82**:5 (1976), 711–712. MR 54 #12561 Zbl 0331.05106

[Bodlaender 1992] H. L. Bodlaender, "On the complexity of some coloring games", pp. 30–40 in *Graph-theoretic concepts in computer science*, edited by R. Möhring, Lecture Notes in Comput. Sci. **484**, Springer, Berlin, 1992. MR 92e:90151 Zbl 0770.90098

[Chou et al. 2003] C.-Y. Chou, W. Wang, and X. Zhu, "Relaxed game chromatic number of graphs", *Discrete Math.* **262**:1-3 (2003), 89–98. MR 2003m:05062 Zbl 1012.05067

[Cowen et al. 1986] L. J. Cowen, R. H. Cowen, and D. R. Woodall, "Defective colorings of graphs in surfaces: partitions into subgraphs of bounded valency", *J. Graph Theory* **10**:2 (1986), 187–195. MR 88c:05056 Zbl 0596.05024

[Cowen et al. 1997] L. Cowen, W. Goddard, and C. E. Jesurum, "Defective coloring revisited", *J. Graph Theory* **24**:3 (1997), 205–219. MR 97m:05091 Zbl 0877.05019

[Dunn and Kierstead 2004a] C. Dunn and H. A. Kierstead, "The relaxed game chromatic number of outerplanar graphs", *J. Graph Theory* **46**:1 (2004), 69–78. MR 2004m:05097 Zbl 1042.05038

[Dunn and Kierstead 2004b] C. Dunn and H. A. Kierstead, "A simple competitive graph coloring algorithm, II", *J. Combin. Theory Ser. B* **90**:1 (2004), 93–106. MR 2005h:05072 Zbl 1033.05039

[Dunn and Kierstead 2004c] C. Dunn and H. A. Kierstead, "A simple competitive graph coloring algorithm, III", *J. Combin. Theory Ser. B* **92**:1 (2004), 137–150. MR 2007b:05069 Zbl 1056.05056

[Eaton and Hull 1999] N. Eaton and T. Hull, "Defective list colorings of planar graphs", *Bull. Inst. Combin. Appl.* **25** (1999), 79–87. MR 99i:05078 Zbl 0916.05026

[Guan and Zhu 1999] D. J. Guan and X. Zhu, "Game chromatic number of outerplanar graphs", *J. Graph Theory* **30**:1 (1999), 67–70. MR 99g:05076 Zbl 0929.05032

[He et al. 2004] W. He, J. Wu, and X. Zhu, "Relaxed game chromatic number of trees and outerplanar graphs", *Discrete Math.* **281**:1-3 (2004), 209–219. MR 2005a:05088 Zbl 1042.05042

Received: 2010-08-27 Revised: 2011-02-10 Accepted: 2011-02-11

chuckl@linfield.edu *Department of Mathematics, Linfield College, 900 SE Baker Street, Unit A468, McMinnville, OR 97128, United States*

jfirkins@linfield.edu *Department of Mathematics, Linfield College, 900 SE Baker Street, Unit A468, McMinnville, OR 97128, United States*

cnaymie@gmail.com *University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada*

epitney@mac.com *Meadow Park Middle School, Beaverton School District, 16550 SW Merlo Road, Beaverton, OR 97006, United States*

william.sehorn@gmail.com *Whitworth University, 300 W. Hawthorne Road, Spokane, WA 99251, United States*

suerchaj@gmail.com *Department of Mathematics, University of Louisville, Louisville, KY 40292, United States*

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.139

Cost-conscious voters in referendum elections

Kyle Golenbiewski, Jonathan K. Hodge and Lisa Moats

(Communicated by Kenneth S. Berenhaut)

In referendum elections, voters are frequently required to register simultaneous yes/no votes on multiple proposals. The separability problem occurs when a voter's preferred outcome on a proposal or set of proposals depends on the known or predicted outcomes of other proposals in the election. Here we investigate cost-consciousness as a potential cause of nonseparability. We develop a mathematical model of cost-consciousness, and we show that this model induces nonseparable preferences in all but the most extreme cases. We show that when outcome costs are distinct, cost-conscious electorates always exhibit both a weak Condorcet winner and a weak Condorcet loser. Finally, we show that preferences consistent with our model of cost-consciousness are rare in randomly generated electorates. We then discuss the implications of our work and suggest directions for further research.

1. Introduction

In referendum elections, voters are often required to register simultaneous yes/no votes on multiple proposals. Recent research demonstrates that the outcomes of such elections can be unsatisfactory or even paradoxical. For example, Lacy and Niou show that the winning outcome can be the last choice of every voter; they argue that this and other troublesome behavior occurs because "referendum elections as currently practiced force people to separate their votes on issues that may be linked in their minds" [Lacy and Niou 2000, page 6].

The phenomenon to which Lacy and Niou allude is known as the *separability problem* [Brams et al. 1997]. What they and others have observed is that voter preferences often contain interdependencies that cannot be expressed through the

MSC2010: 91B12.
Keywords: referendum elections, cost-conscious, separability, separable preferences.
This research was supported by the National Science Foundation under grant no. DMS-0451254, which funds a Research Experiences for Undergraduates (REU) program at Grand Valley State University. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

standard simultaneous method of voting in a referendum. In other words, a voter's preferences on a proposal or a set of proposals may depend on the outcome of another proposal or a set of remaining proposals. Preferences that exhibit this kind of interdependence are said to be *nonseparable*.

Separability has been studied in a variety of contexts, with much of the most recent research focusing on the *structure* and *effects* of separable and nonseparable preferences. Here we take a different approach by investigating one of the underlying *causes* of nonseparability — namely, *cost-consciousness* within the electorate.

To illustrate, consider an election with multiple bond proposals, all competing for funds from the same tax base. In such an election, a voter who is cost-conscious — that is, who desires to limit the total expenditure of public funds — may vote no on a proposal that she supports in principle if she suspects that other proposals are more likely to pass. In doing so, the voter is acting based on predictions about the potential outcomes of these other proposals. If her predictions are wrong, then her voting strategy may also be wrong, or at least less than optimal. In other words, the voter's cost-consciousness complicates the decisions she must make about how to vote on each of the individual proposals. As we will see, these complications can have disastrous effects on the desirability of election outcomes.

Our goal in this paper is to formalize and investigate the consequences of cost-consciousness in referendum elections. Section 2 introduces a model of cost-conscious voter preferences, which we use to show how cost-consciousness induces nonseparability in voter preferences in Section 3. Section 4 demonstrates the existence of Condorcet winning and losing outcomes in certain cost-conscious electorates. Section 5 generalizes the original model by allowing voters to approve of outcomes that exceed their ideal maximum cost, provided that certain conditions are met. Section 6 explores the relative prevalence of cost-conscious voter preferences in randomly generated electorates. Finally, Section 7 summarizes our results and their implications.

2. Model for cost-conscious voters

For the purposes of our investigations, we assume the context of a referendum election on a set Q of $n \geq 2$ questions or proposals. Each potential outcome is represented by an ordered n-tuple of zeros and ones, with 1 typically representing passage of a proposal and 0 representing failure. We let X be the set of all 2^n possible election outcomes. For each $q \in Q$, we let $C(q)$ denote the cost of passing question q, where $C(q) \in \mathbb{R}^+$. The total cost incurred by an election outcome $x \in X$ is then given by

$$C(x) = \sum_{q=1}^{n} x_q C(q),$$

where $x_q = 1$ if question q passes in outcome x, and $x_q = 0$ if question q fails to pass in outcome x. For any subset S of Q, we let $C(S)$ denote the cost of passing all proposals in S; that is,

$$C(S) = \sum_{q \in S} C(q).$$

In general, we assume that each voter's preferences can be represented by a total order on X. This assumption simplifies our analysis and is consistent with prior research on the separability problem in referendum elections. We define a *cost-conscious voter* v to be one who, in principle, supports all of the proposals in Q, but in practice, wishes to limit total spending to some fixed amount M_v.

Definition 2.1. Let v be a voter whose preferences are represented by a total order $\succ$ on X. Then v is said to be *cost-conscious* if there exists some $M_v > 0$ (called the *cost ceiling* for v) such that for each $x, y \in X$, the following axioms hold:

Axiom 1. If $C(x), C(y) \leq M_v$ and $C(y) > C(x)$, then $y \succ x$.

Axiom 2. If $C(x) < C(y)$ and $C(y) > M_v$, then $x \succ y$.

Inherent in Definition 2.1 is the assumption that each voter derives a benefit from each passed proposal that is directly proportional to its cost. In fact, we assume that, for outcomes whose total cost is less than or equal to M_v, the total benefit outweighs the total cost, giving a nonnegative net utility. Furthermore, the utility of each outcome is an increasing function of its cost, provided that the cost does not exceed M_v. Outcomes whose costs exceed M_v have negative net utility, with the net utility decreasing as the cost increases further beyond M_v.

The sudden switch from positive to negative net utility creates a discontinuity in the utility function of each voter at M_v. This discontinuity is reasonable, since M_v marks a cost threshold beyond which outcomes can be thought of as being substantially less attractive, impractical, or even completely unacceptable. For instance, a consumer who has access to $40,000 of credit may attempt to purchase a new car that has as many options as possible, provided that the total cost remains at or below $40,000. Once the $40,000 threshold is exceeded, the consumer may have to go to great lengths in order to purchase the vehicle, if it is even possible for her to do so. In terms of negotiation theory, the $40,000 threshold can be viewed as a *resistance point* — that is, a point beyond which the negotiator would rather do nothing than incur further cost. We postulate that voters can have resistance points for a variety of reasons, both practical and psychological. For instance, a voter may simply be disinclined to approve any package of bond proposals whose total cost exceeds $1 million.

In our initial investigations, we assume that cost ceilings are absolute. That is, they cannot be exceeded without penalty for any reason. In Section 5, we relax this

condition somewhat by allowing voters to exceed their cost ceilings when certain conditions are met.

To illustrate Definition 2.1, suppose

$$|Q| = 3, \qquad C(1) = 200,$$
$$C(2) = 400, \quad C(3) = 500.$$

Furthermore, suppose $M_v = 800$ for some voter v. We note that of the eight possible outcomes, only two have a total cost exceeding M_v— namely,

$$C(1) + C(2) + C(3) = 1100 \quad \text{and} \quad C(2) + C(3) = 900.$$

Thus, Axioms 1 and 2 induce the following ordering on the set of all possible outcomes: $101 \succ 110 \succ 001 \succ 010 \succ 100 \succ 000 \succ 011 \succ 111$. This ordering can also be represented by a preference matrix P_v, as shown below. (For a more detailed treatment of preference matrices, see [Bradley et al. 2005].)

$$P_v = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Now, suppose v becomes more cost-conscious, decreasing M_v to 600. In this case, the outcome 101, which has a cost of 700, is no longer the voter's most preferred outcome. In fact, it becomes the voter's third to last choice. The new induced order is $110 \succ 001 \succ 010 \succ 100 \succ 000 \succ 101 \succ 011 \succ 111$, which corresponds to the preference matrix

$$P_v' = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Note that both P_v and P_v' are uniquely determined by Axioms 1 and 2, once v's cost ceiling and the proposal costs are specified. In particular, the axioms

require each outcome whose cost exceeds M_v to be ranked lower than each outcome whose cost does not exceed M_v. Axiom 1 requires the outcomes whose costs do not exceed M_v to be ranked in descending order with respect to cost, whereas Axiom 2 requires the outcomes whose costs do exceed M_v to be ranked in ascending order with respect to cost. As long as no two outcomes have the same cost, these requirements are enough to induce a unique ordering on X.

Theorem 2.2. *Let v be a cost-conscious voter with cost ceiling M_v, and suppose $C(x) \neq C(y)$ for all distinct $x, y \in X$. Then there is exactly one total order on X that is consistent with Axioms 1 and 2.*

Note that Axioms 1 and 2 impose no restrictions on the ordering of outcomes whose costs are equal. As such, the requirement that no two outcomes have the same cost is essential to Theorem 2.2. To illustrate, consider the case in which $|Q| = 3$, $C(1) = 200$, $C(2) = 300$, and $C(3) = 500$. Since $C(001) = C(110)$, both $001 \succ 110$ or $110 \succ 001$ are permissible by Axioms 1 and 2, regardless of the value of M_v. Thus the conclusion of Theorem 2.2 fails to hold in this case.

3. Cost-consciousness and separability

In Section 1, we suggested that cost-consciousness is a cause of interdependence, or *nonseparability*, within voter preferences. In order to explore this assertion more, we must first define more what it means for a voter's preferences to be separable. Although a more formal treatment of separability can be found in a variety of sources (see, e.g., [Bradley et al. 2005]), the following informal definition will be sufficient for our purposes.

Definition 3.1. Let S be a proper, nonempty subset of Q, and let v be any voter. Then S is said to be *separable* with respect to v if v's preferences over the outcomes of questions within S do not depend on the known or predicted outcomes of questions outside of S.

To illustrate this definition, consider again the preference matrix P_v (Section 2):

$$P_v = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Because $101 \succ 001$, we see that when the outcome on questions 2 and 3 is 01, v prefers 1 to 0 (passage to failure) on question 1. However, if the outcome on

questions 2 and 3 is 11, then v prefers 0 to 1 (failure to passage) on question 1 (since $011 \succ 111$). In other words, voter v's preference on question 1 depends on the outcomes of questions 2 and 3. Because of this, we say that the set $\{1\}$ is nonseparable with respect to v. Note that, from a cost-consciousness standpoint, the nonseparability of $\{1\}$ with respect to v stems from the fact that v wants question 1 to pass if and only if the cost of the other passed proposals in the election is less than or equal to 600.

In contrast, note that regardless of whether question 1 passes or not, voter v always ranks the outcomes of questions 2 and 3 in the same order:

$$01 \succ 10 \succ 00 \succ 11.$$

This is because, for each of these outcomes, the additional passage or failure of question 1 has no bearing on whether the overall cost exceeds v's cost ceiling of 800. Thus for outcomes on $\{2, 3\}$ that cost less than 800 (01, 10, and 00), the more costly outcomes are preferred (by Axiom 1), regardless of whether question 1 passes or not. All of these outcomes are preferred to 11, which always yields a total cost of more than 800 — either with or without the passage of question 1. Because v's ordering of the outcomes on $\{2, 3\}$ does not depend on the outcome of question 1, we say that the set $\{2, 3\}$ is *separable* with respect to v.

The observations from the previous example generalize easily to the following theorem, whose proof is straightforward and thus omitted.

Theorem 3.2. *Let S be a nonempty, proper subset of Q.*

(i) *If $C(Q) > M_v$, then S is separable only if $C(S) > M_v$.*

(ii) *If $C(Q) \leq M_v$, then S is always separable.*

Theorem 3.2 guarantees that the preferences of cost-conscious voters will exhibit some degree of nonseparability, except in two extreme cases. The first is when each proposal, by itself, is more expensive than the voter's cost ceiling. In this case, the voter always prefers failure to passage. The second is when the total cost of all proposals is less than or equal to the voter's cost ceiling. In this case, cost-consciousness is a moot point, and the voter always prefers passage to failure. In every other case, the preferences of cost-conscious voters will exhibit at least some nontrivial interdependencies. The fact that these interdependencies can cause serious problems is illustrated by the following example.

Example 3.3. Consider again an election with $|Q| = 3$, $C(1) = 200$, $C(2) = 400$, and $C(3) = 500$. Suppose that the electorate is comprised of three voters, v_1, v_2, and v_3, for whom $M_{v_1} = 1000$, $M_{v_2} = 800$, and $M_{v_3} = 600$. Then Axioms 1 and 2

uniquely determine the voters' preferences, as follows:

$$P_{v_1} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad P_{v_2} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad P_{v_3} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

With these preferences, each question passes with two "yes" votes and one "no" vote. However, that this outcome (111) is the least preferred choice of every voter. This kind of paradoxical behavior was first observed by Lacy and Niou [2000], but here we have shown it to result from a set of realistic voter preferences — in particular, those consistent with a reasonable model of cost-consciousness.

4. Condorcet winners and losers

In Example 3.3, we saw how a collection of cost-conscious voters could inadvertently elect the worst possible outcome for each voter. It is interesting to note that, in that example, the outcome 101 is a Condorcet winner. The fact that such an outcome exists is not coincidental. In fact, the next theorem establishes that when outcome costs are distinct (as in Theorem 2.2), the assumption of cost-consciousness guarantees the existence of at least a weak Condorcet winner, which we define as follows:

Definition 4.1. Let V be a nonempty collection of voters, and for each $v \in V$, let $\succ_v$ denote a total order on X. An outcome $w \in X$ is said to be a *weak Condorcet winner* (with respect to V) provided that for each $y \in X$ with $y \neq w$,

$$|\{v \in V : w \succ_v y\}| \geq |\{v \in V : y \succ_v w\}|.$$

Theorem 4.2. *Suppose $C(x) \neq C(y)$ for all distinct $x, y \in X$, and let V be any nonempty collection of cost-conscious voters. Then X contains a weak Condorcet winner with respect to V.*

Proof. Let $|Q| = n$. Then X contains 2^n distinct outcomes, which we denote by $x_1, x_2, \ldots, x_{2^n}$. Without loss of generality, assume that

$$C(x_{2^n}) > C(x_{2^n-1}) > \cdots > C(x_2) > C(x_1).$$

Then $x_1 = 00\cdots0$ and $x_{2^n} = 11\cdots1$. We claim that there are 2^n possible preference matrices consistent with Axioms 1 and 2, each determined by the size of M_v in

$$\underset{P_1}{\begin{pmatrix} 11\cdots1 \\ x_{2^n-1} \\ x_{2^n-2} \\ x_{2^n-3} \\ \vdots \\ x_4 \\ x_3 \\ x_2 \\ 00\cdots0 \end{pmatrix}} \quad \underset{P_2}{\begin{pmatrix} x_{2^n-1} \\ x_{2^n-2} \\ x_{2^n-3} \\ x_{2^n-4} \\ \vdots \\ x_3 \\ x_2 \\ 00\cdots0 \\ 11\cdots1 \end{pmatrix}} \quad \underset{P_3}{\begin{pmatrix} x_{2^n-2} \\ x_{2^n-3} \\ x_{2^n-4} \\ x_{2^n-5} \\ \vdots \\ x_2 \\ 00\cdots0 \\ x_{2^n-1} \\ 11\cdots1 \end{pmatrix}} \quad \cdots \quad \underset{P_i}{\begin{pmatrix} x_{2^n-i+1} \\ x_{2^n-i} \\ x_{2^n-i-1} \\ \vdots \\ 00\cdots0 \\ x_{2^n-i+2} \\ x_{2^n-i+3} \\ \vdots \\ 11\cdots1 \end{pmatrix}} \quad \cdots \quad \underset{P_{2^n}}{\begin{pmatrix} 00\cdots0 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_{2^n-3} \\ x_{2^n-2} \\ x_{2^n-1} \\ 11\cdots1 \end{pmatrix}}$$

Table 1. All possible preference matrices for cost-conscious voters, assuming distinct outcome costs.

comparison to the cost of the outcomes in X (see Table 1). In particular, if v is a voter with preference matrix P_v and cost-ceiling M_v, then

$$P_v = P_1 \quad \text{if } M_v \geq C(x_{2^n}),$$
$$P_v = P_2 \quad \text{if } C(x_{2^n}) > M_v \geq C(x_{2^n-1}),$$

and in general,

$$P_v = P_i \quad \text{if } C(x_{2^n-i+2}) > M_v \geq C(x_{2^n-i+1}).$$

Let $|V| = m$, and let m_j denote the number of voters in V with preference matrix P_j. Now suppose that, for some i,

$$\frac{1}{m}\sum_{j=1}^{i-1} m_j < 0.5 \quad \text{and} \quad \frac{1}{m}\sum_{j=1}^{i} m_j \geq 0.5.$$

Then, for each $k = (i+1), (i+2), \ldots, 2^n$, the outcome x_{2^n-i+1} is ranked higher than the outcome x_{2^n-k+1} by at least 50% of voters in V. Also, for each $k = 1, 2, \ldots, (i-2), (i-1)$, the outcome x_{2^n-i+1} is ranked lower than the outcome x_{2^n-k+1} by less than 50% of the voters in V. Since there must be a smallest i for which $(1/m)\sum_{j=1}^{i} m_j \geq 0.5$, the corresponding outcome x_{2^n-i+1} is a weak Condorcet winner with respect to V. □

To illustrate that Theorem 4.2 can fail when two outcomes in X have the same cost, consider the following example:

Example 4.3. Suppose that in an election with three proposals and three voters, $C(1) = C(2) = C(3) = 400$, and $M_v = 500$ for each v. In this case, each voter's preference matrix could be one of 36 distinct options. Suppose that the voters' preference matrices are as follows:

$$P_{v_1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad P_{v_2} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad P_{v_3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Since the outcomes 001, 010, and 100 comprise the top three choices for each voter, any Condorcet winner for this electorate must be one of these three outcomes. However, since the societal preference among these outcomes is cyclic (100 defeats 001, which defeats 010, which defeats 100), there can be no Condorcet winner.

Just as a weak Condorcet winner is guaranteed to exist when outcome costs are distinct, a weak Condorcet loser (defined analogously to Definition 4.1) can also be found in these circumstances.

Theorem 4.4. *Suppose $C(x) \neq C(y)$ for all distinct $x, y \in X$, and let V be any nonempty collection of cost-conscious voters. Then X contains a weak Condorcet loser with respect to V. Furthermore, this weak Condorcet loser is always either $00\cdots0$ or $11\cdots1$.*

Proof. By the same argument as in the proof of Theorem 4.2, each voter's preferences can be represented by one of the 2^n matrices in Table 1. The preference matrix P_1 is the only preference matrix that has the outcome $00\cdots0$ ranked as the least preferred outcome. Every other preference matrix has the outcome $11\cdots1$ ranked as the least preferred outcome. Consider three cases:

Case 1: Less than 50% of voters in V have preference matrix P_1. In this case, more than 50% of voters have preference matrices P_2 through P_{2^n}. Since $11\cdots1$ is the least preferred outcome in P_2 through P_{2^n}, $11\cdots1$ is ranked as the lowest outcome by more than 50% of the voters in V. Thus $11\cdots1$ is a Condorcet loser.

Case 2: Exactly 50% of the voters in V have preference matrix P_1. Then exactly 50% of the voters in V have preference matrices P_2 through P_{2^n}. Since $00\cdots0$ is the least preferred outcome in P_1 and $11\cdots1$ is the least preferred outcome in P_2 through P_{2^n}, $00\cdots0$ is ranked lower than every other outcome by 50% of the voters and $11\cdots1$ is ranked lower than every other outcome by 50% of voters. Thus, both $00\cdots0$ and $11\cdots1$ are weak Condorcet losers.

Case 3: More than 50% of the voters in V have preference matrix P_1. Then, since $00\cdots0$ is the least preferred outcome in P_1, $00\cdots0$ is ranked as the lowest outcome by more than 50% of voters in V. Thus, $00\cdots0$ is a Condorcet loser.

In each case, either $00\cdots0$ or $11\cdots1$ is a weak Condorcet loser, as desired. □

It is worth noting that the proof of Theorem 4.4 depends only on the placement of the outcomes $00\cdots0$ and $11\cdots1$ within the matrices P_1, P_2, ..., P_{2^n}, and not on the relative rankings of other outcomes. Since $00\cdots0$ and $11\cdots1$ will always be the unique least expensive and most expensive outcomes, respectively, the proof would still be valid even without the assumption of distinct outcome costs. Thus, the conclusion of Theorem 4.4 holds even when some of these costs are equal.

5. Weak cost-consciousness

Up to this point, we have assumed that cost-conscious voters are universally resistant to exceeding their cost ceilings. That is, outcomes whose costs exceed M_v are necessarily less preferred than those whose costs do not exceed M_v.

There may, however, be circumstances in which a voter can gain a significant additional benefit by exceeding his or her cost ceiling by a small amount. In this section, we modify our original model of cost-consciousness to allow for such deviations. Our modifications assume that voters are willing to exceed their cost ceiling only when (i) the excess is bounded within a specified tolerance; and (ii) all other options for increasing the voter's total benefit also cause the voter's cost ceiling to be exceeded.

To formulate these conditions more precisely, we must first introduce some new terminology. First, for any outcome $x \in X$, we define the *support set* of x, denoted $S(x)$ to be the set of all questions passed in x. That is,

$$S(x) = \{q \in Q : x_q = 1\}.$$

For all $x, y \in X$, if $S(x) \subset S(y)$, we say that y *augments* x. If $|S(x)| = 1$, then x is said to be a *singleton*. An outcome x is said to be *cost-maximal* if $C(x) \leq M_v$ and there does not exist an outcome $y \in X$ such that y augments x and $C(y) \leq M_v$.

Definition 5.1. Let v be a voter whose preferences are represented by a total order $\succ$ on X. Then v is said to be *weakly cost-conscious* if there exists some $M_v > 0$ (called the *cost ceiling* for v) and some nonnegative $\tau \leq M_v$ (called the *tolerance* for v) such that for each $x, y \in X$, the following axioms hold:

Axiom 1. If $C(x), C(y) \leq M_v$ and $C(y) > C(x)$, then $y \succ x$.

Axiom 2′. If $C(x) < C(y)$ and $C(y) > M_v + \tau$, then $x \succ y$.

Axiom 3. If x is cost-maximal, y augments x, and $M_v < C(y) \leq M_v + \tau$, then $y \succ x$.

Note that when $\tau = 0$, Definition 5.1 is equivalent to Definition 2.1. The next example illustrates the effect of allowing τ to be nonzero.

Example 5.2. Consider an election with three proposals in which $C(1) = 200$, $C(2) = 400$, and $C(3) = 501$. Suppose also that for some voter v, $M_v = 700$ and

$\tau = 0$. Then Theorem 2.2 (which applies since $\tau = 0$ and all outcome costs are unique) guarantees a unique preference matrix consistent with Axioms 1 and 2. In this case, the matrix is

$$P_v = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Note that the outcome 101, with a cost of 701, is the voter's third least preferred outcome. Note, however, that 101 augments three other outcomes: 000, 100, and 001. Of these three outcomes, only the latter is cost-maximal. Thus, if $\tau = 1$, then Axiom 3 requires $101 \succ 001$. This leaves two possibilities for v's now weakly cost-conscious preferences:

$$P_v = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{or} \quad P_v = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Note that the first matrix can be obtained by simply increasing M_v to 701, keeping τ fixed at 0. However, the second matrix cannot be obtained in this way and is in fact inconsistent with our original definition of cost-consciousness. This contrast demonstrates that the flexibility afforded by allowing τ to be nonzero cannot be accomplished by simply increasing M_v.

6. Prevalence of cost-conscious voters

As we showed in Section 3, cost-consciousness can be a significant cause of nonseparability in voter preferences over multiple issues. Hodge and TerHaar [2008] have also shown that the vast majority of randomly selected preference matrices correspond to completely nonseparable preferences — that is, preferences for which every nonempty, proper subset of Q is nonseparable. In light of these observations, it is natural to consider how prevalent cost-conscious preferences are among all possible preference orders. In this section, we will show that the

proportion of total orders on X that are consistent with the axioms of weak cost-consciousness approaches 0 asymptotically. In particular, we will prove the following theorem:

Theorem 6.1. *Let Ω_n denote the set of all total orders on X that are consistent with Axioms 1, 2′, and 3. Then*

$$\lim_{n\to\infty} \frac{|\Omega_n|}{2^n!} = 0.$$

To prove Theorem 6.1 we establish several lemmas, each of which assumes that $\succ$ represents the preferences of a weakly cost-conscious voter. Lemmas 6.2 and 6.3 follow immediately from Axioms 1 and 2′, respectively.

Lemma 6.2. *If $00\cdots0 \succ x$ for some $x \in X$, then $C(x) > M_v$.*

Lemma 6.3. *If $x \succ 00\cdots0$ for some $x \in X$, then $C(x) \leq M_v + \tau$.*

Lemma 6.4. *Let $x, y \in X$ with $S(x) \cap S(y) = \varnothing$. If $x \succ 11\cdots1 \succ 00\cdots0$, then $11\cdots1 \succ y \succ 00\cdots0$.*

Proof. By assumption, there is an outcome $x \in X$ such that $x \succ 11\cdots1$. Consequently, Axiom 1 implies that $C(11\cdots1) > M_v$. Since $11\cdots1 \succ 00\cdots0$, Lemma 6.3 implies that $C(11\cdots1) \leq M_v + \tau$. Since $\tau \leq M_v$, it follows that

$$M_v < C(11\cdots1) \leq M_v + \tau \leq 2M_v.$$

Since $S(x) \cap S(y) = \varnothing$, we know that $C(x) + C(y) \leq C(11\cdots1) \leq 2M_v$. Thus, either $C(x) \leq M_v$ or $C(y) \leq M_v$.

Suppose $C(x) \leq M_v$. Then either x is cost-maximal or there exists a cost-maximal outcome that augments x. To account for either of these cases, let x' denote a cost-maximal element that is either equal to x or augments x. Note that since $C(11\cdots1) > M_v$, $x' \neq 11\cdots1$. Thus, $11\cdots1$ augments x', which implies by Axiom 3 that $11\cdots1 \succ x'$. But since $C(x) \leq C(x') \leq M_v$, Axiom 1 implies that $x' \succeq x$. So $11\cdots1 \succ x' \succeq x$, a contradiction.

Since it cannot be the case that $C(x) \leq M_v$, it must be that $C(y) \leq M_v$. But then an argument similar to that in the preceding paragraph establishes that $11\cdots1 \succ y$. Since $C(y) \leq M_v$, we know also that $y \succ 00\cdots0$ (by Axiom 1. Thus,

$$11\cdots1 \succ y \succ 00\cdots0,$$

as desired. □

Lemma 6.5. *If $00\cdots0 \succ 11\cdots1$, then $C(11\cdots1) > M_v + \tau$.*

Proof. Assume, to the contrary, that $00\cdots0 \succ 11\cdots1$ and $C(11\cdots1) \leq M_v + \tau$. By Lemma 6.2, $M_v < C(11\cdots1)$. Thus, $M_v < C(11\cdots1) \leq M_v + \tau$. Since $\tau \leq M_v$, there exists a cost-maximal $x \in X$ such that $0 < C(x) \leq M_v$. Since

$11\cdots1$ augments x, it follows by Axioms 1 and 3 that $11\cdots1 \succ x \succ 00\cdots0$, a contradiction to the assumption that $00\cdots0 \succ 11\cdots1$. □

Lemma 6.6. *If* $00\cdots0 \succ 11\cdots1$, *then* $x \succ 11\cdots1$ *for all* $x \in X$.

Proof. By Lemma 6.5, $C(11\cdots1) > M_v+\tau$. But for all $x \in X$, $C(x) < C(11\cdots1)$. Therefore, $x \succ 11\cdots1$ by Axiom 2′. □

Lemma 6.7. *If* $11\cdots1 \succ 00\cdots0$, *then there exists* $x \in X$ *such that*

$$11\cdots1 \succ x \succ 00\cdots0.$$

Proof. If $11\cdots1 \succ 00\cdots0$, then $C(11\cdots1) \le M_v + \tau$ by Lemma 6.3. Now consider two cases:

Case 1: If $C(11\cdots1) \le M_v$, then there exists $x \in X$ such that

$$C(00\cdots0) < C(x) < C(11\cdots1) \le M_v.$$

So, by Axiom 1, $11\cdots1 \succ x \succ 00\cdots0$.

Case 2: If $M_v < C(11\cdots1) \le M_v + \tau$, then $\tau \le M_v$ implies that there exists a cost-maximal $x \in X$ such that $0 < C(x) \le M_v$. Since $11\cdots1$ augments x, it follows by Axioms 1 and 3 that $11\cdots1 \succ x \succ 00\cdots0$. □

Lemma 6.7 can be stated more concisely by simply noting that $11\cdots1$ cannot *cover* $00\cdots0$. In general x is said to cover z (with respect to $\succ$) if $x \succ z$ and there does not exist y such that $x \succ y \succ z$.

We are now able to prove Theorem 6.1.

Proof of Theorem 6.1. Let A and B to be the collections of total orders on X defined as follows:

$$A = \{\succ: 00\cdots0 \succ 11\cdots1 \text{ and } 11\cdots1 \succ x \text{ for some } x \in X\}.$$

$$B = \{\succ: 11\cdots1 \text{ covers } 00\cdots0 \text{ with respect to } \succ\}.$$

Furthermore, let C be the collection of total orders $\succ$ on X that satisfy all of the following conditions:

1. $11\cdots1 \succ 00\cdots0$.
2. $11\cdots1$ covers some nonsingleton element z of X, where $z \ne 00\cdots0$.
3. For some singletons $x, y \in X$, either

$$x \succ y \succ 11\cdots1 \quad \text{or} \quad x \succ 11\cdots1 \succ 00\cdots0 \succ y.$$

Note that $A \not\subseteq \Omega_n$, $B \not\subseteq \Omega_n$, and $C \not\subseteq \Omega_n$ by Lemmas 6.6, 6.7, and 6.4, respectively. Note also that A, B, and C are pairwise disjoint. Thus,

$$|\Omega_n| \le 2^n! - |A \cup B \cup C| = 2^n! - |A| - |B| - |C|.$$

It can be easily shown that

$$|A| = \binom{2^n - 1}{2}(2^n - 2)! \quad \text{and} \quad |B| = (2^n - 1)(2^n - 2)!.$$

Thus,

$$\begin{aligned} |A| + |B| &= \frac{(2^n - 1)!}{2!(2^n - 3)!}(2^n - 2)! + (2^n - 1)(2^n - 2)! \\ &= \frac{(2^n - 2)}{2}(2^n - 1)! + (2^n - 1)! \\ &= 2^{n-1}(2^n - 1)!. \end{aligned}$$

To count the elements of C, we note that every order $\succ$ from C can be constructed via a sequence of five choices.

First, we choose z, the nonsingleton element of X that is covered by $11\cdots1$. There are $2^n - n - 2$ possible choices (excluding $11\cdots1$, $00\cdots0$, and the n singleton outcomes).

Next, we divide the singleton elements of X into three groups according to their ranking relative with respect to $11\cdots1$ and $00\cdots0$. In particular, let X' denote the set of singleton elements of X, and let

$$\begin{aligned} i &= |\{x \in X' : x \succ 11\cdots1\}|, \\ j &= |\{x \in X' : 11\cdots1 \succ z \succ x \succ 00\cdots0\}|, \\ k &= |\{x \in X' : 00\cdots0 \succ x\}|. \end{aligned}$$

Note that $i + j + k = n$. Furthermore, the definition of C requires that $i \neq 0$, and if $i = 1$, $k \neq 0$. Any values of i, j, and k that satisfy these conditions will yield a grouping consistent with the definition of C. Thus, there are

$$\binom{n+2}{2} - (n+1) - 1 = \frac{(n+2)(n-1)}{2}$$

such groupings.

Next, we choose an ordering for the n singletons. There are $n!$ such choices.

Our first three steps produce a unique ordering of the singleton elements of X along with the elements $11\cdots1$, z, and $00\cdots0$. Now we must choose which of the 2^n positions in the ranking induced by $\succ$ will be occupied by these $n+3$ outcomes. Since $11\cdots1$ must cover z, we have $\binom{2^n-1}{n+2}$ choices.

Once the positions and ordering of the singletons, $11\cdots1$, z, and $00\cdots0$ are determined, we must choose an ordering for the remaining $2^n - n - 3$ elements of X. There are $(2^n - n - 3)!$ such choices.

Putting all of this together, we obtain:

$$\begin{aligned}|C| &= n!(2^n-n-2)\frac{(n+2)(n-1)}{2}\binom{2^n-1}{n+2}(2^n-n-3)!\\ &= n!(2^n-n-2)\frac{(n+2)(n-1)}{2}\frac{(2^n-1)!}{(n+2)!(2^n-n-3)!}(2^n-n-3)!\\ &= \frac{n!(2^n-n-2)(n+2)(n-1)(2^n-1)!}{2(n+2)!}\\ &= \frac{(n-1)(2^n-n-2)(2^n-1)!}{2(n+1)}.\end{aligned}$$

From this it follows that

$$\begin{aligned}\lim_{n\to\infty}\frac{|\Omega_n|}{2^n!} &\le \lim_{n\to\infty}\frac{2^n!-|A|-|B|-|C|}{2^n!}\\ &= \lim_{n\to\infty}\left(1-\frac{2^{n-1}(2^n-1)!}{2^n!}-\frac{(n-1)(2^n-n-2)(2^n-1)!}{2(n+1)(2^n)!}\right)\\ &= \lim_{n\to\infty}\left(1-\frac{1}{2}-\frac{(n-1)(2^n-n-2)}{2^{n+1}(n+1)}\right)\\ &= \frac{1}{2}-\lim_{n\to\infty}\frac{(n-1)(2^n-n-2)}{2^{n+1}(n+1)} = \frac{1}{2}-\frac{1}{2}=0.\end{aligned}$$

But since $\dfrac{|\Omega_n|}{2^n!}\ge 0$ for all n, it follows that $\lim\limits_{n\to\infty}\dfrac{|\Omega_n|}{2^n!}=0$. □

At first glance, the conclusion of Theorem 6.1 may seem rather surprising. Indeed, one might expect cost-conscious voters to be more prevalent than the theorem suggests. There are a number of reasonable explanations for this apparent discrepancy, all of which warrant further investigation.

First, it may be the case that random samples of preference orders do not accurately represent the preferences of electorates in actual elections. Perhaps some orders are unrealistic and should be eliminated from the start. If this is the case, then among all *realistic* preference orders, however that notion is defined, cost-conscious preferences may be more prevalent. Since random preferences have been used in past research to simulate referendum elections [Hodge and Schwallier 2006], a more careful look at their ability to model actual electorates seems appropriate.

Second, it could be the case that as the number of questions increases, other factors in addition to cost-consciousness have more of an opportunity to play a role in the formation of voter preferences. In other words, while *purely* cost-conscious

preferences may become increasingly rare, the presence of some form of cost-consciousness may still be found, perhaps in abundance.

Finally, our model may not account for all forms of cost-consciousness. In particular, there may be ways of generalizing our model that would allow for a broader range of preferences to be classified as cost-conscious. One direction for further research would be formulate a model based on penalty functions that decrease a voter's net utility in some predictable way when the voter's cost ceiling is exceeded.

7. Summary and conclusions

Cost-consciousness is one cause of nonseparability within voter preferences in multiple-question referendum elections. In fact, cost-consciousness induces preference nonseparability in all but the most trivial of cases. This nonseparability can lead to undesirable election outcomes under the typical method of simultaneous voting.

We have shown that in electorates consisting entirely of cost-conscious voters, a weak Condorcet winner is guaranteed to exist whenever outcome costs are distinct. Furthermore, a weak Condorcet loser is guaranteed to exist whether outcome costs are distinct or not, and this weak Condorcet loser is always either $11\cdots1$ or $00\cdots0$.

Even with a relaxed model of cost-consciousness that allows cost ceilings to be exceeded when certain conditions are met, we showed that preference orders consistent with the axioms of cost-consciousness comprise an arbitrarily small proportion of all possible preferences as the number of questions increases without bound. We discussed several possible explanations for this result, all of which suggest directions for further research.

This research is one of the first attempts to formally model a practical cause of nonseparability in voter preferences over multiple issues. There are certainly other underlying causes of nonseparability, and further investigation of these other causes could eventually lead to the development of a scheme for classifying voter preferences according to the types of interdependence they exhibit.

Our work here has focused on modeling the preferences of cost-conscious voters, but we have not investigated or proposed methods for choosing better election outcomes when electorates are cost-conscious. This direction seems like a natural next step, and one that could potentially have practical implications for the implementation of direct democracy via referendum elections.

References

[Bradley et al. 2005] W. J. Bradley, J. K. Hodge, and D. M. Kilgour, "Separable discrete preferences", *Math. Social Sci.* **49**:3 (2005), 335–353. MR 2005k:91093 Zbl 1114.91030

[Brams et al. 1997] S. J. Brams, D. M. Kilgour, and W. S. Zwicker, "Voting on referenda: The separability problem and possible solutions", *Electoral Studies* **16**:3 (1997), 359–377.

[Hodge and Schwallier 2006] J. K. Hodge and P. Schwallier, "How does separability affect the desirability of referendum election outcomes?", *Theory and Decision* **61**:3 (2006), 251–276. MR 22688736 Zbl 1101.91320

[Hodge and TerHaar 2008] J. K. Hodge and M. TerHaar, "Classifying interdependence in multidimensional binary preferences", *Math. Social Sci.* **55**:2 (2008), 190–204. MR 2009a:91029 Zbl 1143.91012

[Lacy and Niou 2000] D. Lacy and E. M. S. Niou, "A problem with referendums", *Journal of Theoretical Politics* **12**:1 (2000), 5–31.

Received: 2010-09-17 Revised: 2011-02-14 Accepted: 2011-02-16

kyle@math.utk.edu *Department of Mathematics, University of Tennessee, 227 Ayres Hall, 1403 Circle Drive, Knoxville, TN 37996-1320, United States*

hodgejo@gvsu.edu *Department of Mathematics, Grand Valley State University, Allendale, Michigan 49401, United States*

s-lmoats1@math.unl.edu *Department of Mathematics, University of Nebraska, 203 Avery Hall, PO Box 880130, Lincoln, NE 68588-0130, United States*

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.157

On the size of the resonant set for the products of 2×2 matrices

Jeffrey Allen, Benjamin Seeger and Deborah Unger

(Communicated by Chi-Kwong Li)

For $\theta \in [0, 2\pi)$ and $\lambda > 1$, consider the matrix $h = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and the rotation matrix R_θ. Let $W_n(\theta)$ denote some product of m instances of R_θ and n of h, with the condition $m \leq \epsilon n$ $(0 < \epsilon < 1)$. We analyze the measure of the set of θ for which $\|W_n(\theta)\| \geq \lambda^{\delta n}$ $(0 < \delta < 1)$. This can be regarded as a model problem for the Bochi–Fayad conjecture.

1. Introduction

Avila and Roblin [2009] considered the following problem. Take the two matrices

$$H = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \tag{1}$$

and

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta \in [0, 2\pi).$$

Fix $\lambda > 1$ and let $m, n \in \mathbb{N}$. Consider words of the form

$$W_n(\theta) = H^{i_1} R_\theta^{j_1} \dots H^{i_k} R_\theta^{j_k},$$

where k is arbitrary and $i_1, \dots, i_k, j_1, \dots, j_k \in \mathbb{N} \cup \{0\}$ are such that

$$i_1 + \cdots + i_k = n, \quad j_1 + \cdots + j_k = m.$$

Assume that m is much smaller than n and take a "generic" angle θ. It is not unreasonable to conjecture that $\|W_n\|$ grows geometrically with n regardless of the combinatorics of the word. Avila and Roblin proved the following theorem, where the norm is given by $\|W\| = |a| + |b| + |c| + |d|$ if $W = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$.

MSC2010: primary 37H15; secondary 37H05, 37C85.
Keywords: Bochi–Fayad conjecture, resonant set, measure, rotation matrix, Fayad, Krikorian, exponential growth.

Theorem 1. *Assume that $0 < \delta < 1$ is fixed. Then there is an n-independent set Ω such that $|\Omega| = 2\pi$ and for any $\theta \in \Omega$ there is $\epsilon > 0$ so that*

$$\min_{W_n} \|W_n(\theta)\| > \lambda^{\delta n}$$

provided $m < \epsilon n(\ln n \ln\ln n)^{-1}$.

Here the minimum is over all words W_n for n fixed and m as given. This theorem improved earlier results by Fayad and Krikorian [2008]. The special case of the Bochi–Fayad conjecture [Avila and Roblin 2009; Fayad and Krikorian 2008] deals with the similar situation when $m < \epsilon n$ and ϵ is small. One might expect that $|\Omega| \to 2\pi$ as $\epsilon \to 0$ in this case. Proving it seems to be quite hard. We investigate a simpler case. In (1), consider the matrix H when λ is large. Then $\lambda^{-1} \to 0$ as $\lambda \to \infty$ and one might wonder what happens if λ^{-1} is dropped. Thus we consider $h = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & 0 \end{smallmatrix}\right) \notin \mathrm{SL}(2, \mathbb{R})$ instead of H. It turns out that a very precise analysis can be performed for this simpler model problem, as we shall see in the next section. Section 3 provides some numerical evidence and comparison of the model case with the real problem.

2. The model problem

In the previous setting, take $h = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & 0 \end{smallmatrix}\right)$ instead of H and fix $\epsilon \in (0, 1)$. Given n, set

$$f_n(\theta) = \min_{W_n} \|W_n(\theta)\|,$$

where the norm is again given by the sum of absolute values of matrix entries and the minimum is taken over all W_n with $m \le \epsilon n$. Note that we can take the minimum because for a given n there are only finitely many possibilities for W_n.

Finally, we fix $0 < \delta < 1$ and define the *resonant set* $\mathcal{R}$ thus: $\theta \in \mathcal{R}$ if there exists some n such that $f_n(\theta) < \lambda^{\delta n}$. We claim that $|\mathcal{R}| < C\lambda^{-(1-\delta)/\epsilon}$, where C is some constant that can be explicitly computed and $|\mathcal{R}|$ denotes the Lebesgue measure of the set $\mathcal{R}$.

We now make the convention that *there are no zero exponents in the expression of $W_n(\theta)$*. Then, for words having precisely k blocks of rotation matrices, there are four possibilities, differing in which matrix (h or R_θ) begins the word and which matrix ends it:

$$W_n(\theta) = h^{i_1} R_\theta^{j_1} \cdots h^{i_k} R_\theta^{j_k}, \tag{2}$$

$$W_n(\theta) = R_\theta^{j_1} h^{i_1} \cdots R_\theta^{j_k} h^{i_k}, \tag{3}$$

$$W_n(\theta) = R_\theta^{j_1} h^{i_1} \cdots R_\theta^{j_{k-1}} h^{i_{k-1}} R_\theta^{j_k}, \tag{4}$$

$$W_n(\theta) = h^{i_1} R_\theta^{j_1} \cdots h^{i_k} R_\theta^{j_k} h^{i_{k+1}}. \tag{5}$$

For the word in (2), the product has this explicit form:

$$\begin{aligned} W_n(\theta) &= h^{i_1} R_\theta^{j_1} \cdots h^{i_k} R_\theta^{j_k} \\ &= \begin{pmatrix} \lambda^{i_1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \cos j_1\theta & -\sin j_1\theta \\ \sin j_1\theta & \cos j_1\theta \end{pmatrix} \cdots \begin{pmatrix} \lambda^{i_k} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \cos j_k\theta & -\sin j_k\theta \\ \sin j_k\theta & \cos j_k\theta \end{pmatrix} \\ &= \begin{pmatrix} \lambda^{i_1} \cos j_1\theta & -\lambda^{i_1} \sin j_1\theta \\ 0 & 0 \end{pmatrix} \cdots \begin{pmatrix} \lambda^{i_k} \cos j_k\theta & -\lambda^{i_k} \sin j_k\theta \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \lambda^n \cos j_1\theta \cdots \cos j_k\theta & \lambda^n \cos j_1\theta \cdots \cos j_{k-1}\theta \sin j_k\theta \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{6}$$

Likewise, for (3), we obtain

$$\begin{aligned} W_n(\theta) &= R_\theta^{j_1} h^{i_1} \cdots R_\theta^{j_k} h^{i_k} \\ &= \begin{pmatrix} \cos j_1\theta & -\sin j_1\theta \\ \sin j_1\theta & \cos j_1\theta \end{pmatrix} \begin{pmatrix} \lambda^{i_1} & 0 \\ 0 & 0 \end{pmatrix} \cdots \begin{pmatrix} \cos j_k\theta & -\sin j_k\theta \\ \sin j_k\theta & \cos j_k\theta \end{pmatrix} \begin{pmatrix} \lambda^{i_k} & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \lambda^{i_1} \cos j_1\theta & 0 \\ \lambda^{i_1} \sin j_1\theta & 0 \end{pmatrix} \cdots \begin{pmatrix} \lambda^{i_k} \cos j_k\theta & 0 \\ \lambda^{i_k} \sin j_k\theta & 0 \end{pmatrix} \\ &= \begin{pmatrix} \lambda^n \cos j_1\theta \cdots \cos j_k\theta & 0 \\ \lambda^n \sin j_1\theta \cos j_2\theta \cdots \cos j_k\theta & 0 \end{pmatrix}. \end{aligned} \tag{7}$$

Using the result in (7), we get for the word (4)

$$\begin{aligned} &W_n(\theta) \\ &= (R_\theta^{j_1} h^{i_1} \cdots R_\theta^{j_{k-1}} h^{i_{k-1}}) R_\theta^{j_k} \\ &= \begin{pmatrix} \lambda^{i_1+\cdots+i_{k-1}} \cos j_1\theta \cos j_2\theta \cdots \cos j_{k-1}\theta & 0 \\ \lambda^{i_1+\cdots+i_{k-1}} \sin j_1\theta \cos j_2\theta \cdots \cos j_{k-1}\theta & 0 \end{pmatrix} \begin{pmatrix} \cos j_k\theta & -\sin j_k\theta \\ \sin j_k\theta & \cos j_k\theta \end{pmatrix} \\ &= \begin{pmatrix} \lambda^n \cos j_1\theta \cos j_2\theta \cdots \cos j_k\theta & -\lambda^n \cos j_1\theta \cos j_2\theta \cdots \cos j_{k-1}\theta \sin j_k\theta \\ \lambda^n \sin j_1\theta \cos j_2\theta \cdots \cos j_k\theta & -\lambda^n \sin j_1\theta \cos j_2\theta \cdots \cos j_{k-1}\theta \sin j_k\theta \end{pmatrix}. \end{aligned} \tag{8}$$

Finally, using (6), we get for the word (5) simply

$$\begin{aligned} W_n(\theta) &= (h^{i_1} R_\theta^{j_1} \cdots h^{i_{k-1}} R_\theta^{j_{k-1}} h^{i_k} R_\theta^{j_k}) h^{i_{k+1}} \\ &= \begin{pmatrix} \lambda^{i_1+\cdots+i_k} \cos j_1\theta \cdots \cos j_k\theta & \lambda^{i_1+\cdots+i_k} \cos j_1\theta \cdots \cos j_{k-1}\theta \sin j_k\theta \\ 0 & 0 \end{pmatrix} \\ &\qquad \times \begin{pmatrix} \lambda^{i_{k+1}} & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \lambda^n \cos j_1\theta \cdots \cos j_k\theta & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{9}$$

Therefore $\|W_n(\theta)\|$ is given by

$$\begin{aligned}
&\lambda^n|\cos j_1\theta\cdots\cos j_{k-1}\theta|\big(|\cos j_k\theta|+|\sin j_k\theta|\big) && \text{for } W_n \text{ of type (2)},\\
&\lambda^n|\cos j_2\theta\cdots\cos j_k\theta|\big(|\cos j_1\theta|+|\sin j_1\theta|\big) && \text{for } W_n \text{ of type (3)},\\
&\lambda^n|\cos j_2\theta\cdots\cos j_{k-1}\theta|\big(|\cos j_1\theta|+|\sin j_1\theta|\big)\\
&\qquad\qquad\times\big(|\cos j_k\theta|+|\sin j_k\theta|\big) && \text{for } W_n \text{ of type (4)},\\
&\lambda^n|\cos j_1\theta\cdots\cos j_k\theta| && \text{for } W_n \text{ of type (5)}.
\end{aligned}$$

Remark. This shows that, among words with k rotation blocks, $\min\limits_{W_n}\|W_n(\theta)\|$ is reached by words of type (5).

Theorem 2. *Let*

$$\begin{aligned}
S_\alpha &= \big\{\theta\in[0,2\pi)\;\big|\;|\cos\alpha\theta|<\lambda^{-(1-\delta)\alpha/\epsilon-1}\big\},\\
\tilde S_\alpha &= \big\{\theta\in[0,2\pi)\;\big|\;|\cos\alpha\theta|<\lambda^{-(1-\delta)\alpha/\epsilon}\big\}.
\end{aligned}$$

Then the resonant set $\mathcal{R}$ satisfies

$$\bigcup_{\alpha\in\mathbb{N}} S_\alpha\subseteq\mathcal{R}\subseteq\bigcup_{\alpha\in\mathbb{N}}\tilde S_\alpha.$$

Proof. Suppose $\theta\in\bigcup_{\alpha\in\mathbb{N}} S_\alpha$. Then $\theta\in S_\alpha$ for some $\alpha\in\mathbb{N}$ and

$$|\cos\alpha\theta|<\lambda^{-(1-\delta)\alpha/\epsilon-1}.$$

Let $n=[\alpha/\epsilon]+1$. Then, $n-1\le\alpha/\epsilon<n$ and $\alpha<\epsilon n$. Consider the word $\omega_n(\theta)=h^{i_1}R_\theta^\alpha h^{i_2}$ where $i_1+i_2=n$. Since $m=\alpha$, we have $m\le\epsilon n$. Then

$$f_n(\theta)=\min_{W_n(\theta)}\|W_n(\theta)\|\le\|\omega_n(\theta)\|=\lambda^n|\cos\alpha\theta|<\lambda^n\cdot\lambda^{-(1-\delta)\alpha/\epsilon-1}\le\lambda^{\delta n}.$$

Therefore $\theta\in\mathcal{R}$.

Now suppose $\theta\notin\bigcup_{\alpha\in\mathbb{N}}\tilde S_\alpha$. Then $|\cos\alpha\theta|\ge\lambda^{-(1-\delta)\alpha/\epsilon}$ for all $\alpha\in\mathbb{N}$. Choose an arbitrary $n\in\mathbb{N}$. Then

$$\begin{aligned}
f_n(\theta)&=\min_{W_n(\theta)}\|W_n(\theta)\|=\|\omega_n(\theta)\| && \text{(for some word } \omega_n(\theta))\\
&=\lambda^n|\cos j_1\theta\cdots\cos j_k\theta| && \text{(by the remark above)}\\
&=\lambda^n|\cos\alpha_1\theta^{m_1}\cdots\cos\alpha_l\theta^{m_l}|,
\end{aligned}$$

where $\alpha_1<\cdots<\alpha_l$ and $m_1\alpha_1+\cdots+m_l\alpha_l=m\le\epsilon n$. Then

$$\begin{aligned}
f_n(\theta)&=\lambda^n|\cos\alpha_1\theta^{m_1}\cdots\cos\alpha_l\theta^{m_l}|\\
&\ge\lambda^n\cdot\lambda^{-(1-\delta)(m_1\alpha_1+\cdots+m_l\alpha_l)/\epsilon}=\lambda^n\cdot\lambda^{-m(1-\delta)/\epsilon}\ge\lambda^{\delta n},
\end{aligned}$$

and therefore $\theta\notin\mathcal{R}$. □

We claim that $\mathcal{R}$ is a dense open set. To show that $\mathcal{R}$ is open, we show that for each n, f_n is continuous. For each n,

$$R_n = \{\theta \in [0, 2\pi) \mid f_n(\theta) < \lambda^{\delta n}\} = f_n^{-1}((-\infty, \lambda^{\delta n})),$$

which is open as the preimage of a continuous function of an open set. Note that $\mathcal{R} = \bigcup_{n=1}^{\infty} R_n$, a union of open sets, so $\mathcal{R}$ is open.

To show that f_n is continuous, we note that f_n is the minimum of a finite number of continuous functions (the norms of a finite number of words). Denote these functions by $F_1, F_2, \ldots, F_M$, $M \in \mathbb{N}$. Fix arbitrary $\theta \in [0, 2\pi)$, fix $\zeta > 0$, and let $\eta > 0$ be such that whenever $|\theta - \tilde{\theta}| < \eta$, $|F_k(\theta) - F_k(\tilde{\theta})| < \zeta$ for all $k = 1, \ldots, M$. Consider arbitrary $\tilde{\theta} \in (\theta - \eta, \theta + \eta)$. For some i, j,

$$f_n(\theta) = F_i(\theta) \quad \text{and} \quad f_n(\tilde{\theta}) = F_j(\tilde{\theta}).$$

By the definition of f_n,

$$F_i(\theta) \le F_j(\theta) \quad \text{and} \quad F_j(\tilde{\theta}) \le F_i(\tilde{\theta}).$$

Notice that if $F_i(\theta) = F_j(\tilde{\theta})$, then $|f_n(\theta) - f_n(\tilde{\theta})| = 0 < \zeta$ and we are done. Suppose that $F_i(\theta) > F_j(\tilde{\theta})$. Then $|f_n(\theta) - f_n(\tilde{\theta})| = F_i(\theta) - F_j(\tilde{\theta}) \le F_j(\theta) - F_j(\tilde{\theta}) < \zeta$. Otherwise, if $F_i(\theta) < F_j(\tilde{\theta})$, then

$$|f_n(\theta) - f_n(\tilde{\theta})| = F_j(\tilde{\theta}) - F_i(\theta) \le F_i(\tilde{\theta}) - F_i(\theta) < \zeta.$$

To see that $\mathcal{R}$ is dense, let I be any open interval in $[0, 2\pi)$. The collection of points $R_\alpha = \{\pi/2\alpha + (\pi/\alpha)k : k \in \{1, \ldots, 2\alpha - 1\}\}$ is in S_α; indeed, for any $\phi \in R_\alpha$, $\cos \alpha\phi = 0 < \lambda^{-(1-\delta)\alpha/\epsilon - 1}$. If we choose $\alpha > |I|/\pi$, then there must be some element ϕ of R_α in I. Since $\phi \in \bigcup_{\alpha\in\mathbb{N}} S_\alpha \subseteq \mathcal{R} \subseteq \bigcup_{\alpha\in\mathbb{N}} \tilde{S}_\alpha$, we see that every open interval in $[0, 2\pi)$ contains a point in $\mathcal{R}$.

Now we are ready to estimate the size of $\mathcal{R}$. Consider $\tilde{S}_\alpha$ for arbitrary $\alpha \in \mathbb{N}$. The measure of this set is

$$\begin{aligned}
|\tilde{S}_\alpha| &= 4\alpha\Big(\frac{\pi}{2\alpha} - \frac{1}{\alpha}\cos^{-1}(\lambda^{-(1-\delta)\alpha/\epsilon})\Big) \\
&= 2\pi - 4\cos^{-1}(\lambda^{-(1-\delta)\alpha/\epsilon}) \\
&\approx 2\pi - 2\pi + 4\lambda^{-(1-\delta)\alpha/\epsilon} \\
&= 4\lambda^{-(1-\delta)\alpha/\epsilon}.
\end{aligned}$$

Then our estimate for the size of $\mathcal{R}$ is

$$|\mathcal{R}| \le \Big|\bigcup_{\alpha\in\mathbb{N}} \tilde{S}_\alpha\Big| \le \sum_{\alpha=1}^{\infty} |\tilde{S}_\alpha| \approx 4\sum_{\alpha=1}^{\infty} \lambda^{-(1-\delta)\alpha/\epsilon} = \frac{4\lambda^{-(1-\delta)/\epsilon}}{1 - \lambda^{-(1-\delta)/\epsilon}} \approx 4\lambda^{-(1-\delta)/\epsilon},$$

as $\epsilon \sim 0$ and λ, δ are fixed.

3. Some numerical evidence

We provide some numerical and graphical evidence of what was proved. We see how the graphs of the model case and the real case compare for fixed n and m. In addition, it is shown graphically that changing the multiplicities of H affects the word's norm, but in the model case the word's norm is invariant under these changes. The graphs in this section were plotted with Maple 14.

Based on the similarities between the pictures of the real case and the model case, we conjecture that the resonant set in the model case is, in some sense, the limiting set of the resonant set in the real case as λ grows large, since the λ^{-1} term goes to 0 as λ goes to infinity. Of course, since here we take λ relatively small ($\lambda = 2$) for graphing convenience, this is a rough conjecture; in fact, proving it seems to be rather difficult.

Figure 1 shows a red curve and a blue curve. The blue curve is the graph of $\|h^{i_1} R_\theta^2 h^{i_2} R_\theta^3 h^{i_3}\|$ where $i_1, i_2, i_3 \in \mathbb{N}$ and $i_1+i_2+i_3 = 15$. Recall that h is the matrix we use in the model case, where λ^{-1} is replaced by 0. With these combinatorics, varying i_1, i_2, i_3 does not change the graph, as long as their sum is 15. Specifically, Figure 1 is the model case of $n = 15$, $m = 5$ ($j_1 = 2$, $j_2 = 3$), and $\lambda = 2$.

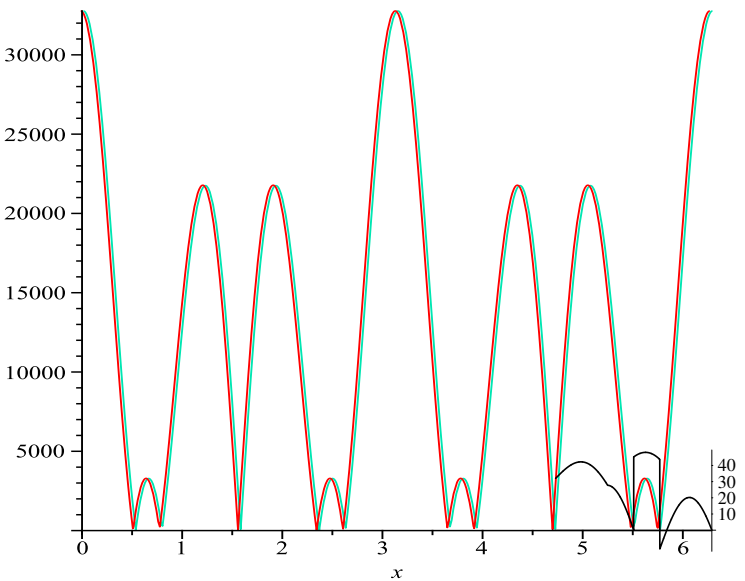


Figure 1. Graphs of the functions $\|H^5 R_\theta^2 H^5 R_\theta^3 H^5\|$ (red curve) and $\|h^5 R_\theta^2 h^5 R_\theta^3 h^5\|$ (blue curve) when $\lambda = 2$. (The curves have been slightly offset horizontally; otherwise they would coincide at this resolution.) The thin black curve on the bottom right is the difference between the first and second functions, with the y-axis expanded 1000 times.

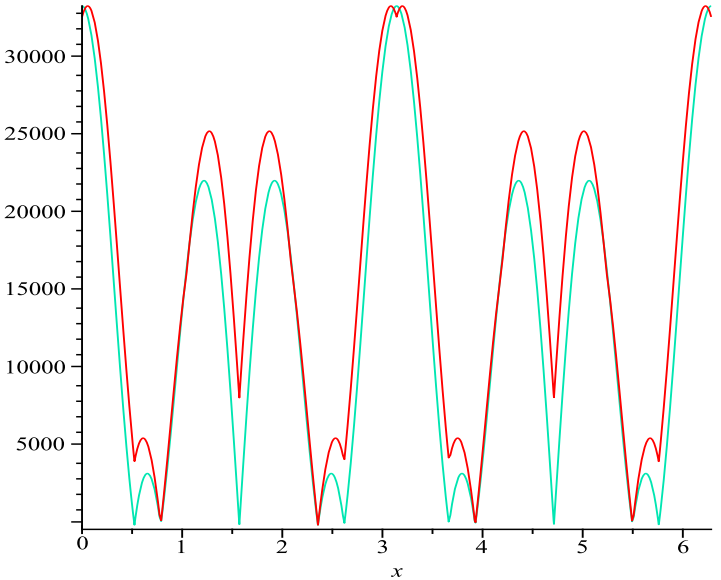


Figure 2. Graphs of the functions $\|H^5R_\theta^2H^9R_\theta^3H^1\|$ (red curve) and $\|h^5R_\theta^2h^9R_\theta^3h^1\|$ (blue curve) when $\lambda = 2$.

The red curve in Figure 1 shows the case where we replace h with H and set $i_1 = i_2 = i_3 = 5$. Explicitly, we are graphing $\|H^5R_\theta^2H^5R_\theta^3H^5\|$.

In Figure 2 we change only two parameters relative to Figure 1: the lengths of the last two blocks of H's (or h's) are now $i_2 = 9$, and $i_3 = 1$. As already seen, the h curve (blue) remains the same, but the H curve (red) — that is, the graph of the function $\|H^5R_\theta^2H^9R_\theta^3H^1\|$ — changes significantly as a result of changing the order of multiplication in the word.

By comparing the red curves in Figures 1 and 2, we observe that a greater disparity between the multiplicities of H (the i_k's) is correlated with a smaller resonant set (the set of points θ between 0 and 2π such that the norm of the word is within a certain distance of zero). The slope of the word's norm is steeper in Figure 2 than it is in Figure 1 and the peaks in Figure 2 are associated with larger values of the word's norm than in the case depicted by Figure 1. Both conditions lead to fewer points θ that are mapped to a norm of the word that is close to zero.

Figure 3 shows the graph of $\|H^1R_\theta^2H^1R_\theta^3H^{13}\|$. Comparing this graph with Figure 2 provides further evidence that a greater disparity between the multiplicities of H results in a smaller resonant set.

To further justify our use of the model case, consider the Figure 4, which treats the case of a word of the form (4). Specifically, the blue curve shows the function $\|R_\theta h^{i_1}R_\theta h^{i_2}R_\theta\|$, where $i_1+i_2 = 15$ and $\lambda = 2$. We take $i_1 = 7$ and $i_2 = 8$ and replace

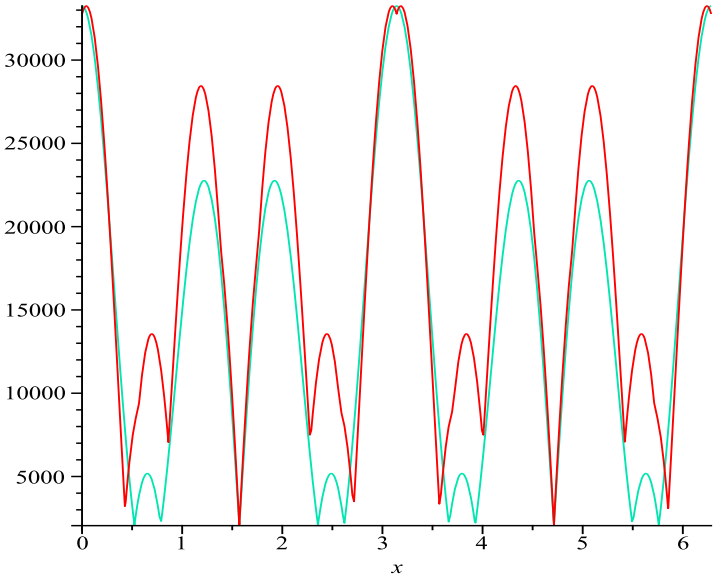


Figure 3. Graphs of the functions $\|H^1 R_\theta^2 H^1 R_\theta^3 H^{13}\|$ (red curve) and $\|h^1 R_\theta^2 h^1 R_\theta^3 h^{13}\|$ (blue curve) when $\lambda = 2$.

h by H to obtain the red curve. As in Figure 1, the two curves are indistinguishable to within the plot's resolution.

Note that the blue curve in Figure 4 is not comparable to that of Figures 1–3. Both show the model case, but with different combinatorics on the word: expression (5) for the earlier figures, and (4) for Figure 4. Both graphs still have a small resonant set.

Figure 5 shows the graphs of $\|R_\theta H^{14} R_\theta H R_\theta\|$ (red) and of $\|R_\theta h^{14} R_\theta h R_\theta\|$ (blue); the latter of course is the same as the blue curve of Figure 4. Comparing Figure 4 with Figure 5, again we see that a greater disparity between the multiplicities of H results in a smaller resonant set.

4. Conclusion

We hope that our model problem is a viable approximation for what happens when the matrix H is used. The next step might be to express the the Bochi–Fayad problem in terms of the model problem. One way to do this might be to write $H = h + e$, where e is the matrix

$$e = \begin{pmatrix} 0 & 0 \\ 0 & \lambda^{-1} \end{pmatrix},$$

and then express a product of H's and R_θ's in terms of a product of h's and R_θ's, and some other, hopefully small, error terms. The numerical evidence above suggests

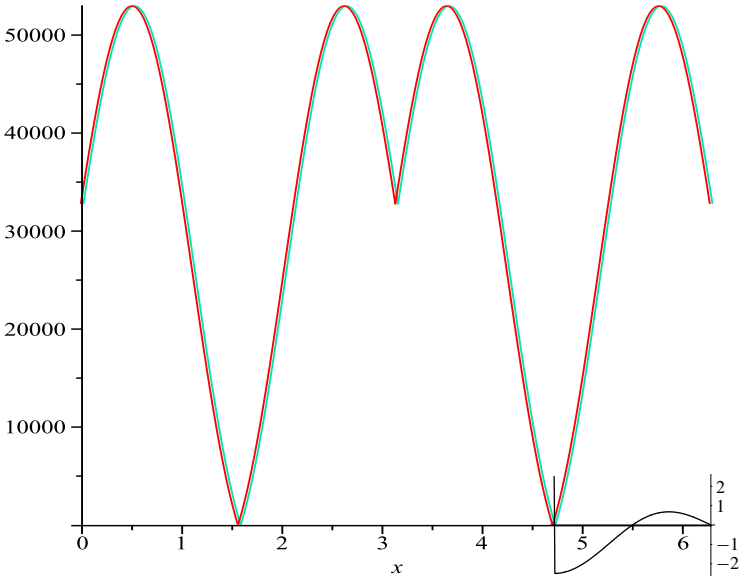


Figure 4. Graphs of the functions $\|R_\theta H^7 R_\theta H^8 R_\theta\|$ (red curve) and $\|R_\theta h^7 R_\theta h^8 R_\theta\|$ (blue curve) when $\lambda = 2$. (The curves have been slightly offset horizontally; otherwise they would coincide at this resolution.) The thin black curve on the bottom right is the difference between the first and second functions, with the y-axis expanded 2500 times.

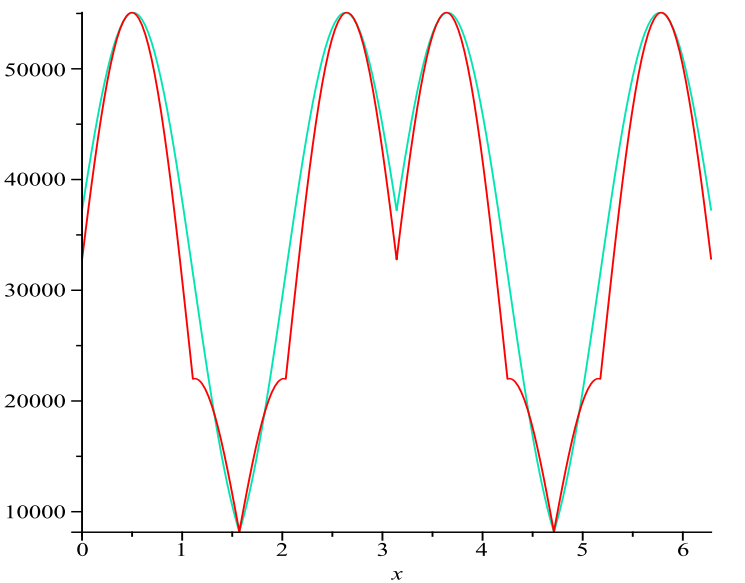


Figure 5. Graphs of $\|R_\theta h^{14} R_\theta h R_\theta\|$ and $\|R_\theta H^{14} R_\theta H R_\theta\|$ when $\lambda = 2$.

that the resonant set for words using the matrix H would actually be smaller than that for words using the matrix h, especially if the distribution of H's is in some sense irregular. This behavior might become more apparent when λ is taken much larger, and ϵ much smaller, than in the experiments above.

Acknowledgements

We would like to thank Professors Serguei Denissov and Alexander Kiselev for their constant support throughout the research and writing processes. We are grateful to them for suggesting the topic and providing us with the opportunity to apply the concepts we have learned throughout our mathematical careers. We thank them both for their excellent advice and encouragement.

References

[Avila and Roblin 2009] A. Avila and T. Roblin, "Uniform exponential growth for some SL(2, ℝ) matrix products", *J. Mod. Dyn.* **3**:4 (2009), 549–554. MR 2011f:37099 Zbl 1189.37060

[Fayad and Krikorian 2008] B. Fayad and R. Krikorian, "Exponential growth of product of matrices in SL(2, ℝ)", *Nonlinearity* **21**:2 (2008), 319–323. MR 2009d:37041 Zbl 1142.37024

Received: 2010-12-10 Revised: 2011-02-17 Accepted: 2011-04-03

jhallen2@wisc.edu *University Of Wisconsin – Madison, Madison, WI 53706, United States*

bseeger@wisc.edu *University Of Wisconsin – Madison, Madison, WI 53706, United States*

dunger@wisc.edu *University Of Wisconsin – Madison, Madison, WI 53715, United States*

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.167

Continuous p-Bessel mappings and continuous p-frames in Banach spaces

Mohammad Hasan Faroughi and Elnaz Osgooei

(Communicated by David R. Larson)

We define the concept of continuous p-frames (cp-frames) for Banach spaces, generalizing discrete p-frames. We prove that under certain conditions the direct sum of a finite number of cp-frames is again a cp-frame. We obtain equivalent conditions for duals of cp-Bessel mappings and show existence and uniqueness of duals of independent cp-frames. Lastly we discuss perturbation of these frames.

1. Introduction

Frames were first introduced in the context of nonharmonic Fourier series [Duffin and Schaeffer 1952]. Outside of signal processing, frames did not seem to generate much interest until the groundbreaking work [Daubechies et al. 1986]. Today, the theory of discrete frames plays an important role not just in digital signal processing and scientific computation, but also in pure and applied mathematics. The interested reader is referred to [Han and Larson 2000; Heil and Walnut 1989] for theory and applications of frames.

A discrete frame is a countable family of elements in a separable Hilbert space which allows stable not necessarily unique decomposition of arbitrary elements into expansions of the frame elements. This concept was generalized in [Ali et al. 1993] to families indexed by some locally compact space endowed with a Radon measure; these frames are known as continuous frames. For more studies about frame theory and continuous frames we refer to [Christensen 2003; Ali et al. 1993; Gabardo and Han 2003; Rahimi et al. 2006].

Various generalizations of frames have been proposed recently, such as frames of subspaces [Asgari and Khosravi 2005], p-frames [Aldroubi et al. 2001; Cao et al. 2008; Christensen and Stoeva 2003], p-frames of subspaces [Najati and Faroughi 2007], g-frames [Sun 2006], and continuous g-frames [Abdollahpour and Faroughi

MSC2010: primary 42C99, 42C15; secondary 42C40.
Keywords: frames, continuous p-frames, Schauder basis, reflexive space.
This is part of Osgooei's Ph.D. thesis at Tabriz University.

2008; Joveini and Amini 2009]. We take as our starting point the generalization presented in [Christensen and Stoeva 2003].

Throughout this paper, (Ω, μ) will be a measure space and μ a positive, σ-finite measure. X is a Banach space with dual X^*. We choose $1 < p < \infty$ and q such that $\frac{1}{p} + \frac{1}{q} = 1$. The normed dual X^* of a Banach space X is itself a Banach space and hence has a normed dual of its own, denoted by X^{**}. A mapping $\Lambda_X : X \mapsto X^{**}$ is well defined by the equation $\langle x, x^* \rangle = \langle x^*, \Lambda_X x \rangle$ for each $x^* \in X^*$; also, $\|\Lambda_X x\| = \|x\|$ for each $x \in X$. So $\Lambda_X : X \to X^{**}$ is an isometric isomorphism of X onto a closed subspace of X^{**}. If X is a reflexive Banach space then Λ_X is an isometric isomorphism of X onto X^{**}.

Definition 1.1. A countable family $\{g_i\}_{i=1}^{\infty} \subset X^*$ is a p-frame for X if there exist constants $A, B > 0$ such that

$$A\|f\| \le \left(\sum_{i=1}^{\infty} |g_i(f)|^p \right)^{1/p} \le B\|f\|. \tag{1-1}$$

If at least the second of these inequalities, called the upper p-frame condition, is satisfied, we say that $\{g_i\}$ is a p-Bessel sequence.

Definition 1.2. Let H be a complex Hilbert space and (Ω, μ) a measure space. A map $F : \Omega \to H$ is called weakly measurable if, for each $f \in H$, the function on Ω defined by $\omega \mapsto \langle f, F(\omega) \rangle$ is measurable. F is called a continuous frame for H with respect to (Ω, μ) if F is weakly measurable and there exist constants $A, B > 0$ such that

$$A\|f\|^2 \le \int_{\Omega} |\langle f, F(\omega) \rangle|^2 d\mu(\omega) \le B\|f\|^2, \quad f \in H. \tag{1-2}$$

In the next results, $R(\cdot)$ denotes the range of a map.

Lemma 1.3 [Rudin 1973]. *Suppose X and Y are Banach spaces and $T \in B(X, Y)$. Then $R(T) = Y$ if and only if $\|T^* y^*\| \ge c\|y^*\|$ for some constant $c > 0$ and for each $y^* \in Y^*$.*

Theorem 1.4 [Rudin 1974]. *$L^p(\Omega, \mu)$ is isometricly isomorphism to the dual space of $L^q(\Omega, \mu)$ via the mapping $K^p : L^p(\Omega, \mu) \to L^q(\Omega, \mu)^*$ give by*

$$K^p \psi(\phi) = \int_{\Omega} \psi(\omega)\phi(\omega)\, d\mu(\omega)$$

for all $\psi \in L^p(\Omega, \mu)$ and $\phi \in L^q(\Omega, \mu)$. We can define an isometric isomorphism $K^q = (K^p)^* \Lambda_q : L^q(\Omega, \mu) \to L^p(\Omega, \mu)^*$ for which Λ_q is the isometric isomorphism of $L^q(\Omega, \mu)$ onto $L^q(\Omega, \mu)^{**}$.

Lemma 1.5 [Heuser 1982]. *Given a bounded operator $U : X \to Y$, the adjoint $U^* : Y^* \to X^*$ is surjective if and only if U has a bounded inverse on $R(U)$.*

Theorem 1.6 [Douglas 1972]. *Let X and Y be Banach spaces. For all $x \in X$ and $y \in Y$, define the 1-norm, $\|(x, y)\|_1 = \|x\|_X + \|y\|_Y$ and the ∞-norm $\|(x, y)\|_\infty = \sup\{\|x\|_X, \|y\|_Y\}$ on the algebraic direct sum $X \oplus Y$. Then $X \oplus Y$ is a Banach space with respect to both norms and these two norms are equivalent.*

In Section 2, we define the concept of cp-Bessel mappings and cp-frames in Banach spaces and show that under some conditions the direct sum of a finite number of cp-frames is again a cp-frame. In Section 3, we define the concept of a cq-Riesz basis and study some relations between cp-frames and cq-Riesz bases. In Section 4, we present a cp-frame mapping $S_F : X \to X^*$ and show that two cp-frames are similar if and only if their analysis operators have the same range. We obtain some equivalent conditions for duals of cp-Bessel mappings and show existence and uniqueness of duals of independent cp-frames in Section 5 and finally in Section 6 we discuss the perturbation of these frames.

2. Continuous p-frames

Definition 2.1. A mapping $F : \Omega \to X^*$ is called a cp-frame for X with respect to (Ω, μ) if F is weakly measurable (Definition 1.2) and there exist positive constants A and B such that

$$A\|x\| \le \left(\int_\Omega |\langle x, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} \le B\|x\|, \quad x \in X. \tag{2-1}$$

The constants A and B are called the lower and upper cp-frame bounds, respectively. F is called a tight cp-frame if A and B can be chosen such that $A = B$, and a Parseval cp-frame if A and B can be chosen such that $A = B = 1$.

F is called a cp-Bessel mapping for X with respect to (Ω, μ) if it is weakly measurable and the second inequality in (2-1) holds. In this case B is called a cp-Bessel constant.

If, in the definition of a cp-frame, we take $\Omega = \mathbb{N}$ and let μ be the counting measure, then our cp-frame will be a p-frame; thus we expect that some properties of p-frames can be satisfied in cp-frames.

Throughout this paper, we simply say F is a cp-frame for X and F is a cp-Bessel mapping for X, instead of F is a cp-frame for X with respect to (Ω, μ) and F is a cp-Bessel mapping for X with respect to (Ω, μ), respectively.

Our study of a cp-frame is based on analysis of two operators,

$$U_F : X \to L^p(\Omega, \mu) \quad \text{and} \quad T_F : L^q(\Omega, \mu) \to X^*.$$

The first is defined by

$$U_F x(\omega) = \langle x, F(\omega)\rangle, \quad x \in X, \quad \omega \in \Omega, \tag{2-2}$$

and the second is weakly defined by

$$T_F\phi(x) = \langle x, T_F\phi\rangle = \int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega), \quad \phi \in L^q(\Omega,\mu),\ x \in X. \tag{2-3}$$

It is clear that if F is a cp-Bessel mapping, then U_F is well defined and bounded operator. U_F is called the analysis and T_F is called the synthesis operator of F.

Lemma 2.2. *Let F be a cp-frame for X. Then the operator $U_F : X \to L^p(\Omega,\mu)$, given by (2-2), has a closed range and X is reflexive.*

Proof. It is easy to verify that U_F has a closed range. By the cp-frame condition, X is isomorphic to $R(U_F)$, but $R(U_F)$ is reflexive because it is a closed subspace of the reflexive space $L^p(\Omega,\mu)$ and therefore X is reflexive. □

Theorem 2.3. *Let $F : \Omega \to X^*$ be a cp-Bessel mapping for X with Bessel bound B. Then the operator $T_F : L^q(\Omega,\mu) \to X^*$, weakly defined in (2-3), is well defined, linear and $\|T_F\| \le B$.*

Proof. It is straightforward. □

Lemma 2.4. *Let $F : \Omega \to X^*$ be a cp-Bessel mapping for X.*

(i) $U_F^* = T_F(K^q)^{-1}$.

(ii) *If X is reflexive, then $T_F^* = K^p U_F \Lambda_X^{-1}$.*

Proof. (i) Since F is a cp-Bessel mapping for X, there exists a unique operator $U_F^* : L^p(\Omega,\mu)^* \to X^*$ such that

$$\langle x, U_F^*\psi\rangle = \langle U_F x, \psi\rangle, \quad x \in X,\ \psi \in L^p(\Omega,\mu)^*.$$

Using Theorem 1.4, we can find $\phi \in L^q(\Omega,\mu)$ such that $K^q(\phi) = \psi$. So, for all $x \in X$ and $\psi \in L^p(\Omega,\mu)^*$,

$$\begin{aligned}\langle x, U_F^*\psi\rangle = \langle U_F x, \psi\rangle = \langle U_F x, K^q(\phi)\rangle &= \int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega)\\ &= \langle x, T_F(\phi)\rangle = \langle x, T_F(K^q)^{-1}\psi\rangle.\end{aligned}$$

Therefore $U_F^* = T_F(K^q)^{-1}$.

(ii) By Theorem 2.3, T_F is well defined and bounded. So for all $f \in X^{**}$ and $\phi \in L^q(\Omega,\mu)$ we have $\langle \phi, T_F^* f\rangle = \langle T_F\phi, f\rangle$. Since X is reflexive, for each $f \in X^{**}$ we can find $x \in X$ such that $\Lambda_X x = f$. Therefore

$$\begin{aligned}\langle \phi, T_F^* f\rangle = \langle T_F\phi, f\rangle = \langle T_F\phi, \Lambda_X x\rangle = \langle x, T_F\phi\rangle &= \int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega)\\ = K^p(\langle x, F\rangle)(\phi) = K^p(\langle \Lambda_X^{-1} f, F\rangle)(\phi) &= \langle \phi, K^p U_F \Lambda_X^{-1} f\rangle.\end{aligned}$$

So $T_F^* = K^p U_F \Lambda_X^{-1}$. □

Theorem 2.5. *Let X be a reflexive Banach space and $F : \Omega \to X^*$ be weakly measurable. If the mapping $T_F : L^q(\Omega, \mu) \to X^*$ weakly defined by*

$$\langle x, T_F\phi\rangle = \int_\Omega \phi(\omega)\langle x, F(\omega)\rangle d\mu(\omega), \quad \phi \in L^q(\Omega, \mu),\ x \in X,$$

is a bounded operator and $\|T_F\| \le B$, then F is a cp*-Bessel mapping for X.*

Proof. Since T_F is well defined and bounded, we have for all $f \in X^{**}$ and $\phi \in L^q(\Omega, \mu)$

$$\langle \phi, T_F^* f\rangle = \langle T_F\phi, f\rangle = \int_\Omega \phi(\omega)\langle \Lambda_X^{-1} f, F(\omega)\rangle\, d\mu(\omega).$$

For each $f \in X^{**}$, we define $\psi_f : \Omega \to \mathbb{C}$ by $\psi_f(\omega) = \langle \Lambda_X^{-1} f, F(\omega)\rangle$. Since ψ_f is measurable and

$$\left|\int_\Omega \phi(\omega)\psi_f(\omega)\, d\mu(\omega)\right| < \infty \quad \text{for all } \phi \in L^q(\Omega, \mu),$$

we obtain $\psi_f \in L^p(\Omega, \mu)$. By Theorem 1.4, we have

$$\psi_f(\omega) = (K^p)^{-1}(T_F^* f)(\omega), \quad \omega \in \Omega.$$

Hence, for each $x \in X$,

$$\left(\int_\Omega |\langle x, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} = \|(K^p)^{-1} T_F^* \Lambda_X x\| = \|T_F^* \Lambda_X x\|$$
$$\le \|T_F^*\|\|x\| \le B\|x\|. \qquad \square$$

Theorem 2.6. *Let X be a reflexive Banach space and $F : \Omega \to X^*$ be a weakly measurable mapping. Then F is a* cp*-frame for X if and only if T_F is a well defined and bounded operator of $L^q(\Omega, \mu)$ onto X^*. In this case, the frame bounds are $\|(T_F^*)^{-1}\|^{-1}$ and $\|T_F\|$.*

Proof. By Theorems 2.3 and 2.5, the upper cp-frame condition satisfies if and only if T_F is well defined and bounded operator of $L^q(\Omega, \mu)$ into X^*. Now suppose that F is a cp-frame for X. Then U_F has a bounded inverse on its range $R(U_F)$ and by Lemma 1.5, U_F^* is surjective and therefore T_F is surjective by Lemma 2.4.

Conversely, suppose that T_F is a well defined and bounded operator of $L^q(\Omega, \mu)$ onto X^*. By Lemma 2.4, for each $x \in X$,

$$\|U_F x\| = \|(K^P)^{-1} T_F^* \Lambda_X x\| = \|T_F^* \Lambda_X x\| \le \|T_F\|\|x\|.$$

On the other hand since T_F is bounded and surjective, T_F^* is one to one, hence T_F^* has a bounded inverse on $R(T_F^*)$. So, by Lemma 2.4, for each $x \in X$ we have

$$\|x\| = \|\Lambda_X x\| = \|(T_F^*)^{-1} T_F^* \Lambda_X x\| \le \|(T_F^*)^{-1}\|\|U_F x\|. \qquad \square$$

Corollary 2.7. *Let $G : \Omega \to X^{**}$ be a weakly measurable mapping. Then the following assertions are equivalent*:

(i) *There exist positive constants A and B such that*

$$A\|g\| \leq \left(\int_\Omega |\langle g, G(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} \leq B\|g\|, \quad g \in X^*.$$

(ii) *X is reflexive and $T_G : L^q(\Omega, \mu) \to X^{**}$ is a well defined, bounded operator of $L^q(\Omega, \mu)$ onto X^{**}.*

Proof. (i) means that $G : \Omega \to X^{**}$ constitutes a cp-frame for X^*. Therefore X^* is reflexive by Lemma 2.2, and thus X is reflexive. The converse is evident by Theorem 2.6. □

Theorem 2.8. *Let X and Y be reflexive Banach spaces. Suppose that $F : \Omega \to X^*$ is a cp-Bessel mapping for X and $W : Y \to X$ is a bounded operator.*

(i) *$W^*F : \Omega \to Y^*$ is a cp-Bessel mapping for Y and $W^*T_F = T_{W^*F}$.*

(ii) *Let $F : \Omega \to X^*$ be a cp-frame for X. Then, W^*F is a cp-frame for Y if and only if W^* is surjective.*

Proof. (i) For each $y \in Y$, the function $\omega \mapsto \langle y, W^*F(\omega)\rangle = \langle Wy, F(\omega)\rangle$ is measurable. Let B be an upper frame bound for F. Then, for each $y \in Y$,

$$\left(\int_\Omega |\langle y, W^*F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} = \left(\int_\Omega |\langle Wy, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} \leq B\|Wy\| \leq B\|W\|\|y\|.$$

Therefore W^*F is a cp-Bessel mapping for Y. For all $y \in Y$ and $\phi \in L^q(\Omega, \mu)$,

$$\langle y, T_{W^*F}\phi\rangle = \int_\Omega \phi(\omega)\langle y, W^*F(\omega)\rangle\, d\mu(\omega) = \int_\Omega \phi(\omega)\langle Wy, F(\omega)\rangle\, d\mu(\omega) = \langle Wy, T_F\phi\rangle = \langle y, W^*T_F\phi\rangle.$$

(ii) If W^* is surjective, then by Theorem 2.6, W^*T_F is surjective. So W^*F is a cp-frame for Y. Conversely, if W^*F is a cp-frame for Y then T_{W^*F} is surjective and so W^* is surjective. □

Proposition 2.9 [Fabian et al. 2001]. *Let Y be a closed subspace of a Banach space Z. If Y is complemented and X is a complement of Y in Z, then Z/Y is isomorphic to X. The dual Z^* is then isomorphic to $Y^* \oplus X^*$; in short, $(Y \oplus X)^* = Y^* \oplus X^*$.*

Theorem 2.10. *Let X and Y be reflexive Banach spaces. Suppose that $F : \Omega \to X^*$ and $G : \Omega \to Y^*$ are cp-Bessel mappings. Then $\psi : \Omega \to X^* \oplus Y^* \cong (X \oplus Y)^*$, $\psi(\omega) = (F(\omega), G(\omega))$ is a cp-Bessel mapping for $X \oplus Y$. The mapping*

$$T_\psi : L^q(\Omega, \mu) \to (X \oplus Y)^* \cong X^* \oplus Y^*$$

is well defined and bounded, and $T_\psi \phi = (T_F \phi, T_G \phi)$ *for all* $\phi \in L^q(\Omega, \mu)$. *Also,*

$$T_\psi^* : (X \oplus Y)^{**} \cong X^{**} \oplus Y^{**} \to L^q(\Omega, \mu)^*$$

is well defined, linear and bounded and $T_\psi^*(f, g) = T_F^* f + T_G^* g$ *for all* (f, g) *in* $X^{**} \oplus Y^{**}$.

Proof. Using Theorem 1.6 and Proposition 2.9, the proof is evident. □

Theorem 2.11. *Let X and Y be reflexive Banach spaces. Suppose that* $F : \Omega \to X^*$ *and* $G : \Omega \to Y^*$ *are* cp*-frames for X and Y, respectively. If* $R(T_F^*) \cap R(T_G^*) = 0$ *and* $R(T_F^*) + R(T_G^*)$ *is a closed subspace of* $L^q(\Omega, \mu)^*$, *then* $\psi : \Omega \to (X \oplus Y)^*$ *is a* cp*-frame for* $X \oplus Y$.

Proof. We define $L : R(T_F^*) \oplus R(T_G^*) \to R(T_F^*) + R(T_G^*)$ by $L(\eta, \gamma) = \eta + \gamma$. Clearly L is well defined, linear and bijective. We have $\|L(\eta, \gamma)\| = \|\eta + \gamma\| \leq (\|\eta\| + \|\gamma\|) = \|(\eta, \gamma)\|_1$. By Theorem 1.6, L is continuous. By the open mapping theorem, L^{-1} is well defined and bounded, since $R(T_F^*) + R(T_G^*)$ is a closed subspace of $L^q(\Omega, \mu)^*$. Therefore by Theorem 1.6, there exists $M > 0$ such that

$$\|(\eta, \gamma)\|_\infty \leq M \|\eta + \gamma\|. \tag{2-4}$$

Let A_1 and A_2 be lower cp-frame bounds for F and G, and set $K = \min\{A_1, A_2\}$. By Theorem 1.6, there exists $M_1 > 0$ such that, for all $(x, y) \in X \oplus Y$,

$$\begin{aligned} K^p \|(x, y)\|_\infty^p &\leq K^p M_1^p (\|x\| + \|y\|)^p \leq K^p M_1^p 2^p (\|x\|^p + \|y\|^p) \\ &\leq 2^p M_1^p \int_\Omega |\langle x, F(\omega)\rangle|^p d\mu(\omega) + 2^p M_1^p \int_\Omega |\langle y, G(\omega)\rangle|^p d\mu(\omega) \\ &\leq 2^p M_1^p \|(K^p)^{-1} T_F^* \Lambda_X x\| + 2^p M_1^p \|(K^p)^{-1} T_G^* \Lambda_Y y\| \\ &= 2^p M_1^p \|T_F^* \Lambda_X x\| + 2^p M_1^p \|T_G^* \Lambda_Y y\| \\ &= 2^p M_1^p \|(T_F^* \Lambda_X x, T_G^* \Lambda_Y y)\|_1, \end{aligned} \tag{2-5}$$

where $\Lambda_X : X \to X^{**}$ and $\Lambda_Y : Y \to Y^{**}$ are isometric isomorphisms of X onto X^{**} and of Y onto Y^{**}, respectively. Again by using Theorem 1.6, there is $M_2 > 0$ such that

$$\|(T_F^* \Lambda_X x, T_G^* \Lambda_Y y)\|_1 \leq M_2 \|(T_F^* \Lambda_X x, T_G^* \Lambda_Y y)\|_\infty. \tag{2-6}$$

By (2-4), (2-5) and (2-6)

$$\begin{aligned} K^p \|(x, y)\|_\infty^p &\leq 2^p M_1^p M_2 M \|T_F^* \Lambda_X x + T_G^* \Lambda_Y y\| = 2^p M_1^p M_2 M \|T_\psi^*(\Lambda_X x, \Lambda_Y y)\| \\ &= 2^p M_1^p M_2 M \|(K^p)^{-1} T_\psi^*(\Lambda_X x, \Lambda_Y y)\| \\ &= 2^p M_1^p M_2 M \|(K^p)^{-1} T_\psi^* \Lambda_{X \oplus Y}(x, y)\| \\ &= 2^p M_1^p M_2 M \int_\Omega |\langle (x, y), \psi(\omega)\rangle|^p d\mu(\omega). \end{aligned}$$

□

Corollary 2.12. *Let $X_1, \cdots, X_n$ be reflexive Banach spaces. Suppose that $F_i : \Omega \to X_i^*$, are* c*p-frames for X_i for all $i \in \mathbb{N}$. If $R(T_{F_j}^*) \cap \left(\sum_{i=1_{i\neq j}}^n R(T_{F_i}^*)\right) = 0$ for each $j \in \mathbb{N}$ and $\sum_{i=1}^n R(T_{F_i}^*)$ is a closed subspace of $L^q(\Omega, \mu)^*$, then the map $\eta : \Omega \to \left(\bigoplus_{i=1}^n X_i\right)^*$ defined by $\eta(\omega) = (F_1(\omega), \cdots, F_n(\omega))$ is a* c*p-frame for $\bigoplus_{i=1}^n X_i$.*

3. Continuous q-Riesz bases

Throughout this paper X is a reflexive Banach space.

Definition 3.1. Let $1 < q < \infty$. A mapping $F : \Omega \to X^*$ is called a cq-Riesz basis for X^* if

(i) $\{x : \langle x, F(\omega)\rangle = 0, w \in \Omega\} = \{0\}$,

(ii) F is weakly measurable, and

(iii) the operator $T_F : L^q(\Omega, \mu) \to X^*$ weakly defined by

$$\langle x, T_F\phi\rangle = \int_\Omega \phi(\omega)\langle x, F(\omega)\rangle \, d\mu(\omega), \quad x \in X, \quad \phi \in L^q(\Omega, \mu),$$

is well defined and there are positive constants A and B such that

$$A\|\phi\|_q \leq \|T_F\phi\|_{X^*} \leq B\|\phi\|_q, \quad \phi \in L^q(\Omega, \mu).$$

A and B are called, respectively, the lower and upper cq-Riesz basis bounds of F.

Theorem 3.2. *Let $F : \Omega \to X^*$ be a* c*q-Riesz basis for X^* with* c*q-Riesz basis bounds A and B. Then F is a* c*p-frame for X with* c*p-frame bounds A and B.*

Proof. Since F is a cq-Riesz basis for X^*, the operator T_F is well defined, bounded and surjective. By Theorem 2.6, F is a cp-frame for X. The upper cq-Riesz basis bound coincide with the upper cp-frame bound by Theorem 2.5. The analogue statement for the lower bound follows from [Dunford and Schwartz 1958, p. 479] and Theorem 2.6. □

Theorem 3.3. *Let $F : \Omega \to X^*$ be a* c*p-frame for X. Then the following statements are equivalent*:

(i) *F is a* c*q-Riesz basis for X^*.*

(ii) *T_F is injective.*

(iii) $R(U_F) = L^p(\Omega, \mu)$.

Proof. (i) $\Longrightarrow$ (ii) By the definition of cq-Riesz basis the proof is evident.

(ii) $\Longrightarrow$ (i) T_F is well defined, bounded and onto by Theorem 2.6, and is injective by (ii), so it has a bounded inverse. Therefore F is a cq-Riesz basis for X^*.

(i) $\Longrightarrow$ (iii) By assumption, T_F has a bounded inverse on $R(T_F) = X^*$. By Lemma 1.5, T_F^* is surjective and Lemma 2.4, implies that $R(U_F) = L^p(\Omega, \mu)$.

(iii) $\Longrightarrow$ (i) is clear. □

4. Maps of cp-frames and their invertibility

In this section, we need a mapping from the Banach space $L^p(\Omega, \mu)$ into its dual space, $L^q(\Omega, \mu)$. For this we use the concept of duality mapping.

First recall that a Banach space X is said to be:

- strictly convex if, whenever $x, y \in X$ with $x \neq y$, $\|x\| = \|y\| = 1$, then $\|\lambda x + (1-\lambda)y\| < 1$ for $\lambda \in (0, 1)$;
- uniformly convex if the conditions $\{x_i\} \subseteq X$, $\{y_i\} \subseteq X$, $\|x_i\| \leq 1$, $\|y_i\| \leq 1$, $\lim_{i\to\infty} \|x_i + y_i\| = 2$, imply that $\lim_{i\to\infty} \|x_i - y_i\| = 0$.

Definition 4.1. The mapping ϕ_X of X into the set of subsets of X^*, defined by

$$\phi_X x = \{x^* \in X^* : x^*(x) = \|x\|\|x^*\|, \|x^*\| = \|x\|\}$$

is called the duality mapping on X.

By the Hahn–Banach theorem $\phi_X x$ is nonempty for all $x \in X$ and $\phi_X 0 = 0$. In general the duality mapping is set-valued, but for certain spaces it is single-valued and such spaces are called smooth.

Proposition 4.2 [Dragomir 2004]. (i) *If X^* is strictly convex then for each $x \in X$, $\phi_X x$ consists of unique element $x^* \in X^*$.*

(ii) *If X and X^* are strictly convex and X is reflexive then ϕ_X is bijective.*

(iii) *If H is a Hilbert space then $\phi_H x = x$ for each $x \in H$.*

Remark 4.3. We can deduce by [Carothers 2005, Corollary 11.13] and [Martin 1976, p. 12] that $L^q(\Omega, \mu)$ is strictly convex.

The next statement is clear from the definition of duality mapping on $L^p(\Omega, \mu)$:

Proposition 4.4. *For all nonzero $\psi \in L^p(\Omega, \mu)$ we have* $\phi_{L^p(\Omega,\mu)}\psi = \dfrac{\overline{\psi}|\psi|^{p-2}}{\|\psi\|_p^{p-2}}$.

Definition 4.5. Let $F : \Omega \to X^*$ be a cp-frame for X. The bounded mapping $S_F : X \to X^*$ defined by $S_F = T_F(K^q)^{-1}\phi_{L^p(\Omega,\mu)}U_F$ will be called a cp-frame mapping of F.

Proposition 4.6. *Suppose that $F : \Omega \to X^*$ is a cp-frame for X with frame bounds A and B. Then S_F has the following properties*:

(i) $S_F = U_F^*\phi_{L^p(\Omega,\mu)}U_F$.

(ii) $A^2\|x\|^2 \leq S_F x(x) \leq B^2\|x\|^2$, $x \in X$.

Proof. Clear from the definition of S_F and of the duality mapping on $L^p(\Omega, \mu)$. □

Definition 4.7. A mapping $[\cdot, \cdot]$ from $X \times X$ into $\mathbb{R}$ is said to be a semi-inner product on X if it has these properties:

(i) $[x, x] \geq 0$ for all $x \in X$ and $[x, x] = 0$ if and only if $x = 0$.

(ii) $[\alpha x + \beta y, z] = \alpha[x, z] + \beta[y, z]$ for all $\alpha, \beta \in \mathbb{R}$ and for all $x, y, z \in X$.

(iii) $|[x, y]|^2 \leq [x, x][y, y]$ for all $x, y \in X$.

If X^* is strictly convex, then there is a unique semi-inner product on X such that $\|x\|_X = [x, x]^{1/2}$ for all $x \in X$ and $\phi_X x(y) = [y, x]$ for all $x, y \in X$ [Dragomir 2004], where ϕ_X is the duality mapping on X. In this case an operator $A : X \to X$ is said to be adjoint abelian if $[Ax, y] = [x, Ay]$ for all $x, y \in X$ or equivalently $A^*\phi_X = \phi_X A$ [Stampfli 1969].

An element $x \in X$ is called (Giles-)orthogonal to $y \in X$, and we write $x \perp y$, if $[y, x] = 0$. If M is a linear subspace of X, the orthogonal complement of M in the Giles sense is denoted by $M^\perp = \{x \in X;\ x \perp y,\ y \in M\}$.

Remark 4.8. Let $F : \Omega \to X^*$ be a cp-frame for X. Suppose that $\mathrm{Ker}(T_F)$ and $(\mathrm{Ker}(T_F))^\perp$ are topologically complementary in $L^q(\Omega, \mu)$, then clearly the operator $T_F|_{(\mathrm{Ker}(T_F))^\perp}$ is invertible and $T_F^\perp = (T_F|_{(\mathrm{Ker}(T_F))^\perp})^{-1}$ is a bounded right inverse of T_F.

Definition 4.9. Let $F : \Omega \to X^*$ be a cp-frame for X. Suppose that $\mathrm{Ker}(T_F)$ and $(\mathrm{Ker}(T_F))^\perp$ are topologically complementary in $L^q(\Omega, \mu)$, we define the mapping $K : X^* \to X$ by $K = \Lambda_X^{-1}(T_F^\perp)^* \phi_{L^q(\Omega,\mu)} T_F^\perp$.

Lemma 4.10. *Let $F : \Omega \to X^*$ be a* cp*-frame for X. Suppose that* $\mathrm{Ker}(T_F)$ *and* $(\mathrm{Ker}(T_F))^\perp$ *are topologically complementary in* $L^q(\Omega, \mu)$.

(i) $K(g)(g) \geq \|g\|_{X^*}^2 / B^2$, *where B denotes an upper* cp*-frame bound for F.*

Moreover, when the operator $T_F^\perp T_F$ is adjoint abelian, the following assertions hold:

(ii) S_F *is invertible and* $S_F^{-1} = K$.

(iii) $S_F^{-1} = U_F^{-1}(K^p)^{-1} \phi_{L^q(\Omega,\mu)} T_F^\perp$.

Proof. The proof is similar to that of [Stoeva 2008, Theorem 5.1]. □

Definition 4.11. Two cp-frames $F : \Omega \to X^*$ and $G : \Omega \to X^*$ for X are similar if there exists an invertible operator $V : X \to X$ such that $F(\omega) = V^* G(\omega)$ for each $\omega \in \Omega$.

Theorem 4.12. *Let the assumptions in Definition 4.9 be satisfied for $F : \Omega \to X^*$ and $G : \Omega \to X^*$. Suppose that $T_F^\perp T_F$ and $T_G^\perp T_G$ are adjoint abelian operators. Then F and G are similar if and only if their analysis operators have same ranges.*

Proof. Suppose F and G are similar. Then there exists an invertible operator $V : X \to X$ such that $F(\omega) = V^*G(\omega)$, $\omega \in \Omega$. Let $\phi \in R(U_F)$. Then there exists $x \in X$, such that

$$\phi(\omega) = U_F x(\omega) = \langle x, F(\omega)\rangle = \langle x, V^*G(\omega)\rangle = U_G(Vx)(\omega), \quad \omega \in \Omega.$$

So $\phi \in R(U_G)$. By a similar argument, $R(U_G) \subseteq R(U_F)$.

Conversely, assume $R(U_F) = R(U_G)$. For each $x \in X$, there is $y \in X$ such that $U_F(x) = U_G(y)$ or $\langle x, F(\omega)\rangle = \langle y, G(\omega)\rangle$, $\omega \in \Omega$. We define the operator $V : X \to X$ by $Vx = y$. Since the cp-frame mappings for F and G are invertible, y is uniquely determined by V and V is linear, one to one and surjective. □

5. Duals of cp-Bessel mappings

In this section, X is an infinite-dimensional, reflexive Banach space.

Definition 5.1 [Fabian et al. 2001]. A sequence $\{e_i\}_{i=1}^{\infty}$ in X is called a Schauder basis of X, if for each $x \in X$ there is a unique sequence of scalars $(a_i)_{i=1}^{\infty}$, called the coordinates of x, such that $x = \sum_{i=1}^{\infty} a_i e_i$.

Let $\{e_i\}_{i=1}^{\infty}$ be a Schauder basis of a Banach space X. For $j \in \mathbb{N}$ and $x = \sum_{i=1}^{\infty} a_i e_i$, denote $f_j(x) = a_j$. Using [Fabian et al. 2001, Theorem 6.5], $f_j \in X^*$. The functionals $\{f_i\}_{i=1}^{\infty}$ are called the associated biorthogonal functionals (coordinate functionals) to $\{e_i\}_{i=1}^{\infty}$ and for each $x \in X$, we have $x = \sum_{i=1}^{\infty} f_i(x)e_i$.

We will denote the biorthogonal functionals $\{f_i\}$ by $\{e_i^*\}$, and say that $\{e_i, e_i^*\}$ is a Schauder basis of X. Such a Schauder basis is called shrinking if $\overline{\text{span}}\{e_i^*\} = X^*$. It is called boundedly complete if $\sum_{i=1}^{\infty} a_i e_i$ converges whenever the scalars a_i are such that $\sup_n \| \sum_{i=1}^{n} a_i e_i\| < \infty$.

Theorem 5.2 [Fabian et al. 2001]. *Let $\{e_i, e_i^*\}$ be a Schauder basis of a Banach space X with the canonical projections $p_n : X \to X$, $p_n(\sum_{i=1}^{\infty} a_i e_i) = \sum_{i=1}^{n} a_i e_i$ for each $n \in \mathbb{N}$. Then the following assertions are equivalent*:

(i) *$\{e_i, e_i^*\}$ is shrinking.*

(ii) *$\{e_i^*, e_i\}$ is a Schauder basis of X^*.*

Theorem 5.3 [Fabian et al. 2001]. *Let X be a Banach space with a Schauder basis $\{e_i, e_i^*\}_{i=1}^{\infty}$. Then X is reflexive if and only if $\{e_i, e_i^*\}$ is both shrinking and boundedly complete.*

Theorem 5.4. *Let $F : \Omega \to X^*$ be a cp-Bessel mapping for X and $G : \Omega \to X^{**}$ be a cq-Bessel mapping for X^*. Then the following assertions are equivalent*:

(i) *For each $x \in X$, $x = \Lambda_X^{-1} T_G (K^p)^{-1} T_F^* \Lambda_X x$.*

(ii) *For each $g \in X^*$, $g = T_F (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g$.*

(iii) *For each $x \in X$ and $g \in X^*$, $\langle x, g\rangle = \int_\Omega \langle x, F(\omega)\rangle \langle g, G(\omega)\rangle \, d\mu(\omega)$.*

(iv) *For each Schauder basis* $\{e_i, e_i^*\}$ *of* X,

$$\langle e_i, e_j^* \rangle = \int_\Omega \langle e_i, F(\omega)\rangle \langle e_j^*, G(\omega)\rangle \, d\mu(\omega), \quad i, j \in \mathbb{N}.$$

Proof. (i) $\Longrightarrow$ (ii) Let $x \in X$ and $g \in X^*$. We have

$$\begin{aligned}
\langle x, g\rangle &= \langle \Lambda_X^{-1} T_G (K^p)^{-1} T_F^* \Lambda_X x, g\rangle = \langle T_G (K^p)^{-1} T_F^* \Lambda_X x, (\Lambda_X^*)^{-1} g\rangle \\
&= \langle (K^p)^{-1} T_F^* \Lambda_X x, T_G^* (\Lambda_X^*)^{-1} g\rangle = \langle T_F^* \Lambda_X x, \Lambda_q (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle \\
&= \langle \Lambda_X x, T_F^{**} \Lambda_q (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle \\
&= \langle \Lambda_X x, (\Lambda_X^{-1})^* T_F (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle \\
&= \langle x, T_F (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle.
\end{aligned}$$

So, for each $g \in X^*$,

$$g = T_F (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g.$$

(ii) $\Longrightarrow$ (iii) For all $x \in X$ and $g \in X^*$,

$$\begin{aligned}
\langle x, g\rangle &= \langle x, T_F (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle \\
&= \int_\Omega \langle x, F(\omega)\rangle (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g(\omega) \, d\mu(\omega). \qquad (5\text{-}1)
\end{aligned}$$

But for all $\psi \in L^p(\Omega, \mu)$ and $h \in X^{***}$ (the dual of X^{**}),

$$\langle \psi, T_G^* h\rangle = \langle T_G \psi, h\rangle = \int_\Omega \psi(\omega) \langle \Lambda_X^* h, G(\omega)\rangle \, d\mu(\omega) = K^q(\langle \Lambda_X^* h, G\rangle)(\psi).$$

So

$$T_G^* h = K^q(\langle \Lambda_X^* h, G\rangle). \qquad (5\text{-}2)$$

Therefore, by (5-1) and (5-2), we have

$$\begin{aligned}
\langle x, g\rangle &= \int_\Omega \langle x, F(\omega)\rangle (K^q)^{-1} K^q(\langle \Lambda_X^* (\Lambda_X^*)^{-1} g, G(\omega)\rangle) \, d\mu(\omega) \\
&= \int_\Omega \langle x, F(\omega)\rangle \langle g, G(\omega)\rangle \, d\mu(\omega).
\end{aligned}$$

(iii) $\Longrightarrow$ (ii) This is clear from the proof of (ii) $\Longrightarrow$ (iii).

(ii) $\Longrightarrow$ (i) For all $x \in X$ and $g \in X^*$, we have

$$\begin{aligned}
\langle x, g\rangle &= \langle x, T_F (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle = \langle\, x, \Lambda_X^* T_F^{**} \Lambda_q (K^q)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle \\
&= \langle T_F^* (\Lambda_X x), \Lambda_q (\Lambda_q)^{-1} ((K^p)^*)^{-1} T_G^* (\Lambda_X^*)^{-1} g\rangle \\
&= \langle T_G (K^p)^{-1} T_F^* (\Lambda_X x), (\Lambda_X^*)^{-1} g\rangle = \langle \Lambda_X^{-1} T_G (K^p)^{-1} T_F^* (\Lambda_X x), g\rangle.
\end{aligned}$$

Since X^* separates the points of X, we get

$$x = \Lambda_X^{-1} T_G (K^p)^{-1} T_F^* (\Lambda_X x), \qquad x \in X.$$

(iii) $\Longrightarrow$ (iv) is obvious.

(iv) $\Longrightarrow$ (iii) For all $x \in X$ and $g \in X^*$,

$$\int_\Omega \langle x, F(\omega)\rangle \langle g, G(\omega)\rangle \, d\mu(\omega) = K^p(\langle x, F\rangle)(\langle g, G\rangle). \tag{5-3}$$

By Theorem 5.2 and 5.3, $\{e_i^*, e_i\}$ and $\{\Lambda e_i, e_i^*\}$ are Schauder basis of X^* and X^{**}, respectively. Therefore

$$\begin{aligned}
K^p(\langle x, F\rangle)(\langle g, G\rangle) &= K^p\left(\left\langle x, \sum_{i=1}^{\infty} \langle e_i, F\rangle e_i^* \right\rangle\right)\left(\left\langle g, \sum_{j=1}^{\infty} \langle e_j^*, G\rangle \Lambda_X e_j \right\rangle\right) \\
&= \left(\sum_{i,j=1}^{\infty} \langle x, e_i^*\rangle \langle g, \Lambda_X e_j\rangle \right) K^p(\langle e_i, F\rangle)(\langle e_j^*, G\rangle) \\
&= \left(\sum_{i,j=1}^{\infty} \langle x, e_i^*\rangle \langle g, \Lambda_X e_j\rangle \right) \int_\Omega \langle e_i, F(\omega)\rangle \langle e_j^*, G(\omega)\rangle \, d\mu(\omega) \\
&= \sum_{i,j=1}^{\infty} \langle x, e_i^*\rangle \langle e_j, g\rangle \langle e_i, e_j^*\rangle \\
&= \left\langle \sum_{i=1}^{\infty} \langle x, e_i^*\rangle e_i, \sum_{j=1}^{\infty} \langle e_j, g\rangle e_j^* \right\rangle = \langle x, g\rangle.
\end{aligned}$$

So, by (5-3),

$$\int_\Omega \langle x, F(\omega)\rangle \langle g, G(\omega)\rangle \, d\mu(\omega) = \langle x, g\rangle. \qquad \square$$

Definition 5.5. Let $F : \Omega \to X^*$ be a cp-Bessel mapping for X and $G : \Omega \to X^{**}$ be a cq-Bessel mapping for X^*. We say that (F, G) is a c-dual pair if one of the assertions of Theorem 5.4 is satisfied.

In this case F is called a cp-dual of G and by Theorem 5.4, we can say that G is a cq-dual of F.

Theorem 5.6. *Let (F, G) be a c-dual pair. Then F is a* cp*-frame for X and G is a* cq*-frame for X^*.*

Proof. For each $x \in X$, we have

$$\begin{aligned}
\|x\| &= \|\Lambda_X^{-1} T_G (K^p)^{-1} T_F^* \Lambda_X x\| = \|T_G (K^p)^{-1} T_F^* \Lambda_X x\| \\
&\le \|T_G\| \|(K^p)^{-1} T_F^* \Lambda_X x\| = \|T_G\| \int_\Omega |\langle x, F(\omega)\rangle|^p d\mu(\omega).
\end{aligned}$$

Since (F, G) is a c-dual pair, $\|T_G\|$ is nonzero. Thus

$$\frac{\|x\|}{\|T_G\|} \leq \left(\int_\Omega |\langle x, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p}.$$

Hence F is a cp-frame for X. We prove similarly that G is a cq-frame for X^*. □

Definition 5.7. Let $F : \Omega \to X^*$ be a cp-frame for X. We say that F is independent if, for every measurable function $\phi : \Omega \to \mathbb{C}$ and every $x \in X$, the condition

$$\int_\Omega \langle x, F(\omega)\rangle \phi(\omega)\, d\mu(\omega) = 0$$

implies that $\phi = 0$.

Theorem 5.8. *Let $F : \Omega \to X^*$ be a cp-frame for X and $\mu(E) \geq k > 0$ for each measurable set E except $E = \varnothing$.*

(i) *If F is an independent cp-frame for X, there exists a unique cq-frame, $G : \Omega \to X^{**}$ for X^*, such that (F, G) is a c-dual pair.*

(ii) *If $\mathrm{Ker}(T_F)$ and $(\mathrm{Ker}(T_F))^\perp$ are topologically complementary in $L^q(\Omega, \mu)$, then there exists a cq-frame $G : \Omega \to X^{**}$ for X^*, such that (F, G) is a c-dual pair.*

Proof. (i) Let F be an independent cp-frame for X. Then $T_F : L^q(\Omega, \mu) \to X^*$ is invertible. We define $G(\omega) = p(\omega)(T_F)^{-1}$, $w \in \Omega$, where $p(\omega) : L^q(\Omega, \mu) \to \mathbb{C}$, defined by $p(\omega)(\phi) = \phi(\omega)$. Now we show that for a fix $\omega_0 \in \Omega$, $p(\omega_0)$ is bounded.

For each $\phi \in L^q(\Omega, \mu)$, $\|\phi\| \leq 1$, put $\Delta = \{\omega \in \Omega : |\phi(\omega)| \geq |\phi(\omega_0)|\}$. Clearly Δ is nonempty and measurable. Since

$$\|\phi\|^q = \int_\Omega |\phi(\omega)|^q\, d\mu(\omega) \geq \int_\Delta |\phi(\omega)|^q\, d\mu(\omega) \geq \mu(\Delta)|\phi(\omega_0)|^q \geq k|\phi(\omega_0)|^q,$$

and

$$\|p(\omega_0)\| = \sup_{\|\phi\|\leq 1} |p(\omega_0)(\phi)| = \sup_{\|\phi\|\leq 1} |\phi(\omega_0)| \leq \sup_{\|\phi\|\leq 1} \left(\frac{1}{k}\right)^{1/q} \|\phi\| = \left(\frac{1}{k}\right)^{1/q},$$

for each $\omega \in \Omega$, $p(\omega)$ is bounded. Therefore $G(\omega) \in X^{**}$. By the definition of $G(\omega)$, for each $g \in X^*$, the mapping $\omega \to \langle g, G(\omega)\rangle$ is measurable and

$$\frac{\|g\|}{\|T_F\|} \leq \left(\int_\Omega |\langle g, G(\omega)\rangle|^q d\mu(\omega)\right)^{1/q} = \|(T_F)^{-1}g\| \leq \|(T_F)^{-1}\|\|g\|.$$

Therefore, G is a cq-frame for X^* with bounds $\|T_F\|^{-1}$ and $\|(T_F)^{-1}\|$.

By the definition of G, $T_G^* = K^q T_F^{-1} \Lambda_X^*$. So, for each $g \in X^*$, we have $g = T_F T_F^{-1}(g) = T_F(K^q)^{-1} T_G^*(\Lambda_X^*)^{-1} g$. Therefore (F, G) is a c-dual pair by Theorem 5.4.

Now we will show the uniqueness of G. Let (F, W) be another c-dual pair. Then

$$T_F(K^q)^{-1}T_G^*(\Lambda_X^*)^{-1} = T_F(K^q)^{-1}T_W^*(\Lambda_X^*)^{-1} = I_{X^*}.$$

Thus $T_G^* = T_W^*$. So $W = G$.

(ii) Since $R(T_F) = X^*$, by Remark 4.8, there is an operator $T_F^{\perp} : X^* \to L^q(\Omega, \mu)$ such that $T_F T_F^{\perp} = I_{X^*}$. For each $g \in X^*$, let $\phi = T_F^{\perp} g$. Therefore for all $x \in X$ and $g \in X^*$

$$\langle x, g\rangle = \langle x, T_F\phi\rangle = \int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega) = \int_\Omega T_F^{\perp} g(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega).$$

For each $\omega \in \Omega$, define $G(\omega) : X^* \to \mathbb{C}$, $G(\omega)(g) = (T_F^{\perp} g)(\omega)$. Then

$$|G(\omega)g| = |p(\omega)(T_F^{\perp} g)| \le \left(\frac{1}{k}\right)^{1/q} \|T_F^{\perp}\|\|g\|,$$

where $p(\omega)$ is defined in the proof of (i). Therefore G is weakly measurable and $G(\omega) \in X^{**}$. Since $T_F T_F^{\perp} = I_{X^*}$, we have, for each $g \in X^*$,

$$\frac{\|g\|}{\|T_F\|} \le \left(\int_\Omega |\langle g, G(\omega)\rangle|^q d\mu(\omega)\right)^{1/q} = \|T_F^{\perp} g\|_q \le \|T_F^{\perp}\|\|g\|. \qquad \square$$

Theorem 5.9. *Let $F : \Omega \to X^*$ be an independent* cp*-frame for X. Suppose that $\mu(E) \ge k > 0$ for each measurable set E except $E = \varnothing$. Let $\omega_0 \in \Omega$ be such that*

$$\mu(\{\omega_0\}) \ne \frac{1}{\langle F(\omega_o), G(\omega_o)\rangle},$$

where $G : \Omega \to X^{**}$ is the unique cq*-dual of F, obtained in Theorem 5.8. Then $F : \Omega \setminus \{\omega_0\} \to X^*$ is a* cp*-frame for X.*

Proof. It is clear that the upper frame condition holds. For the lower frame bound, we have

$$\langle x, F(\omega_0)\rangle = \int_\Omega \langle x, F(\omega)\rangle\langle F(\omega_0), G(\omega)\rangle\, d\mu(\omega), \quad x \in X.$$

Therefore $\langle x, F(\omega_0)\rangle$ is given by

$$\int_{\Omega\setminus\{\omega_0\}} \langle x, F(\omega)\rangle\langle F(\omega_0), G(\omega)\rangle\, d\mu(\omega) + \langle x, F(\omega_0)\rangle\langle F(\omega_0), G(\omega_0)\rangle\mu(\{\omega_0\}),$$

that is,

$$\langle x, F(\omega_0)\rangle = \frac{1}{1 - \mu(\{\omega_0\})\langle F(\omega_0), G(\omega_0)\rangle} \int_{\Omega\setminus\{\omega_0\}} \langle x, F(\omega)\rangle\langle F(\omega_0), G(\omega)\rangle\, d\mu(\omega).$$

Let A be the lower frame bound of F. For each $x \in X$,

$$|\langle x, F(\omega_0)\rangle|^p \le K \int_{\Omega\setminus\{\omega_o\}} |\langle x, F(\omega)\rangle|^p d\mu(\omega),$$

where

$$K = \left(\frac{1}{1-\mu(\{\omega_0\})\langle F(\omega_0), G(\omega_0)\rangle}\right)^p \left(\int_{\Omega\setminus\{\omega_0\}} |\langle F(\omega_0), G(\omega)\rangle|^q d\mu(\omega)\right)^{p/q}.$$

Therefore, for each $x \in X$,

$$\begin{aligned} A\|x\|_X &\le \left(\int_{\Omega\setminus\{\omega_o\}} |\langle x, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} + \left(|\langle x, F(\omega_0)\rangle|^p \mu(\{\omega_0\})\right)^{1/p} \\ &\le \left(\int_{\Omega\setminus\{\omega_o\}} |\langle x, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} \\ &\qquad + \left(\int_{\Omega\setminus\{\omega_o\}} |\langle x, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p} K^{1/p}(\mu(\{\omega_o\}))^{1/p} \\ &= \left(1 + K^{1/p}(\mu(\{\omega_o\}))^{1/p}\right)\left(\int_{\Omega\setminus\{\omega_o\}} |\langle x, F(\omega)\rangle|^p d\mu(\omega)\right)^{1/p}. \end{aligned}$$

Therefore $F : \Omega \setminus \{\omega_0\} \to X^*$ is a cp-frame for X with lower frame bound

$$\frac{A}{1 + K^{1/p}(\mu(\{\omega_o\}))^{1/p}}.$$

□

Corollary 5.10. *Let $F : \Omega \to X^*$ be a* cp*-frame for X and assume $\mu(E) \ge k > 0$ for each measurable set E except $E = \varnothing$. Let $\omega_0 \in \Omega$ be such that*

$$\mu(\{\omega_0\}) \ne \frac{1}{\langle F(\omega_o), G(\omega_o)\rangle}.$$

Suppose $\mathrm{Ker}(T_F)$ *and* $(\mathrm{Ker}(T_F))^\perp$ *are topologically complementary in $L^q(\Omega, \mu)$. Then $F : \Omega \setminus \{\omega_0\} \to X^*$ is a* cp*-frame for X.*

6. Perturbation of cp-frames

Perturbation of discrete frames has been discussed in [Cazassa and Christensen 1997]. The proof of the following theorem is based on the following lemma, which was proved in [Cazassa and Christensen 1997].

Lemma 6.1. *Let U be a linear operator on a Banach space X and assume that there exist $\lambda_1, \lambda_2 \in [0, 1)$ such that for each $x \in X$,*

$$\|x - Ux\| \le \lambda_1\|x\| + \lambda_2\|Ux\|.$$

Then U is bounded and invertible. Moreover, for each $x \in X$,

$$\frac{1-\lambda_1}{1+\lambda_2}\|x\| \le \|Ux\| \le \frac{1+\lambda_1}{1-\lambda_2}\|x\|$$

and

$$\frac{1-\lambda_2}{1+\lambda_1}\|x\| \le \|U^{-1}x\| \le \frac{1+\lambda_2}{1-\lambda_1}\|x\|.$$

Theorem 6.2. *Let F be an independent* cp*-frame for X and $\mu(E) \ge k > 0$, for each measurable set E, except $E = \varnothing$. Suppose that $G : \Omega \to X^*$ is weakly measurable and assume that there exist constants $\lambda_1, \lambda_2, \gamma \ge 0$ with $\max(\lambda_1 + \gamma/A, \lambda_2) < 1$. Suppose also that, for all $\phi \in L^q(\Omega, \mu)$ and x in the unit sphere of X,*

$$\left|\int_\Omega \phi(\omega)\langle x, F(\omega) - G(\omega)\rangle\, d\mu(\omega)\right| \le \lambda_1\left|\int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega)\right| + \lambda_2\left|\int_\Omega \phi(\omega)\langle x, G(\omega)\rangle\, d\mu(\omega)\right| + \gamma\|\phi\|.$$

Then $G : \Omega \to X^*$ is a cp*-frame for X with bounds*

$$A\,\frac{1-(\lambda_1+\gamma/A)}{1+\lambda_2} \quad \text{and} \quad B\,\frac{1+\lambda_1+\gamma/B}{1-\lambda_2},$$

where A and B are the frame bounds of F.

Proof. Let $X_1 = \{x \in X : \|x\| = 1\}$ be the unit sphere of X. We first prove that G is a cp-Bessel mapping for X. By assumption, for all $x \in X$ and $\phi \in L^q(\Omega, \mu)$,

$$\begin{aligned}
&\left|\int_\Omega \phi(\omega)\langle x, G(\omega)\rangle\, d\mu(\omega)\right| \\
&\le \left|\int_\Omega \phi(\omega)\langle x, F(\omega) - G(\omega)\rangle\, d\mu(\omega)\right| + \left|\int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega)\right| \\
&\le (1+\lambda_1)\left|\int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega)\right| + \lambda_2\left|\int_\Omega \phi(\omega)\langle x, G(\omega)\rangle\, d\mu(\omega)\right| + \gamma\|\phi\|,
\end{aligned}$$

which implies that

$$\begin{aligned}
\left|\int_\Omega \phi(\omega)\langle x, G(\omega)\rangle\, d\mu(\omega)\right| &\le \frac{1+\lambda_1}{1-\lambda_2}\left|\int_\Omega \phi(\omega)\langle x, F(\omega)\rangle\, d\mu(\omega)\right| + \frac{\gamma}{1-\lambda_2}\|\phi\| \\
&\le \left(\frac{1+\lambda_1}{1-\lambda_2}B + \frac{\gamma}{1-\lambda_2}\right)\|\phi\|.
\end{aligned}$$

Let $K : L^q(\Omega, \mu) \to X^*$ be defined by

$$\langle x, K\phi\rangle = \int_\Omega \phi(\omega)\langle x, G(\omega)\rangle\, d\mu(\omega), \quad x \in X, \phi \in L^q(\Omega, \mu).$$

Then

$$\|K\phi\| = \sup_{\|x\|=1} |\langle x, K\phi\rangle| = \sup_{\|x\|=1} \left| \int_\Omega \phi(\omega)\langle x, G(\omega)\rangle \, d\mu(\omega) \right|$$
$$\leq \left(\frac{1+\lambda_1}{1-\lambda_2} B + \frac{\gamma}{1-\lambda_2} \right) \|\phi\|.$$

Therefore K is well defined and bounded. So by Theorem 2.5, G is a cp-Bessel mapping for X with upper bound $B(1+\lambda_1+\gamma/B)/(1-\lambda_2)$.

We define $V = K(K^q)^{-1} T_W^* (\Lambda_X^*)^{-1}$, for which W is the unique cq-dual of F which is obtained in Theorem 5.8. Then, for all $x \in X$ and $g \in X^*$,

$$\langle x, Vg\rangle = \langle x, K(K^q)^{-1} T_W^* (\Lambda_X^*)^{-1} g\rangle = \int_\Omega \langle g, W(\omega)\rangle \langle x, G(\omega)\rangle \, d\mu(\omega)$$

and

$$\langle x, g\rangle = \int_\Omega \langle x, F(\omega)\rangle \langle g, W(\omega)\rangle \, d\mu(\omega).$$

Let $\phi_g : \Omega \to \mathbb{C}$ be defined by $\phi_g(\omega) = \langle g, W(\omega)\rangle$. Clearly $\phi_g \in L^q(\Omega, \mu)$. Therefore, by assumption, we deduce that for all $x \in X_1$ and $g \in X^*$,

$$|\langle x, g - Vg\rangle| \leq \lambda_1 |\langle x, g\rangle| + \lambda_2 |\langle x, Vg\rangle| + \gamma \|\phi_g\|.$$

Hence

$$\|g - Vg\| = \sup_{\|x\|=1} |\langle x, g - Vg\rangle| \leq \lambda_1 \|g\| + \lambda_2 \|Vg\| + \gamma \|\phi_g\|$$
$$\leq \left(\lambda_1 + \frac{\gamma}{A}\right) \|g\| + \lambda_2 \|Vg\|.$$

By Lemma 6.1, V is invertible and

$$\|V\| \leq \frac{1+\lambda_1+\gamma/A}{1-\lambda_2}, \quad \|V^{-1}\| \leq \frac{1+\lambda_2}{1-(\lambda_1+\gamma/A)}.$$

Then

$$\langle x, g\rangle = \langle x, VV^{-1} g\rangle = \int_\Omega \langle V^{-1} g, W(\omega)\rangle \langle x, G(\omega)\rangle \, d\mu(\omega),$$

and we obtain

$$\|x\| = \|\Lambda_X x\| = \sup_{\|g\|=1} |\langle g, \Lambda_X x\rangle| = \sup_{\|g\|=1} |\langle x, g\rangle|$$
$$= \sup_{\|g\|=1} \left| \int_\Omega \langle V^{-1} g, W(\omega)\rangle \langle x, G(\omega)\rangle \, d\mu(\omega) \right|$$
$$\leq \sup_{\|g\|=1} \left(\int_\Omega |\langle V^{-1} g, W(\omega)\rangle|^q d\mu(\omega) \right)^{1/q} \left(\int_\Omega |\langle x, G(\omega)\rangle|^p d\mu(\omega) \right)^{1/p}.$$

Therefore, for each $x \in X$,

$$A\,\frac{1-(\lambda_1+\gamma/A)}{1+\lambda_2}\,\|x\| \le \left(\int_\Omega |\langle x, G(\omega)\rangle|^p d\mu(\omega)\right)^{1/p}. \qquad \square$$

Acknowlegement

The authors would like to thank the referee for useful comments and suggestions.

References

[Abdollahpour and Faroughi 2008] M. R. Abdollahpour and M. H. Faroughi, "Continuous G-frames in Hilbert spaces", *Southeast Asian Bull. Math.* **32**:1 (2008), 1–19. MR 2008m:41028 Zbl 1199.42132

[Aldroubi et al. 2001] A. Aldroubi, Q. Sun, and W.-S. Tang, "p-frames and shift invariant subspaces of L^p", *J. Fourier Anal. Appl.* **7**:1 (2001), 1–21. MR 2002c:42046 Zbl 0983.46027

[Ali et al. 1993] S. T. Ali, J.-P. Antoine, and J.-P. Gazeau, "Continuous frames in Hilbert space", *Ann. Physics* **222**:1 (1993), 1–37. MR 94e:81107 Zbl 0782.47019

[Asgari and Khosravi 2005] M. S. Asgari and A. Khosravi, "Frames and bases of subspaces in Hilbert spaces", *J. Math. Anal. Appl.* **308**:2 (2005), 541–553. MR 2006b:42042 Zbl 1091.46006

[Cao et al. 2008] H.-X. Cao, L. Li, Q.-J. Chen, and G.-X. Ji, "(p, Y)-operator frames for a Banach space", *J. Math. Anal. Appl.* **347**:2 (2008), 583–591. MR 2009h:46024 Zbl 05344335

[Carothers 2005] N. L. Carothers, *A short course on Banach space theory*, London Math. Soc. Student Texts **64**, Cambridge University Press, Cambridge, 2005. MR 2005k:46001 Zbl 1072.46001

[Cazassa and Christensen 1997] P. G. Cazassa and O. Christensen, "Perturbation of operators and applications to frame theory", *J. Fourier Anal. Appl.* **3**:5 (1997), 543–557. MR 98j:47028 Zbl 0895.47007

[Christensen 2003] O. Christensen, *An introduction to frames and Riesz bases*, Birkhäuser, Boston, 2003. MR 2003k:42001 Zbl 1017.42022

[Christensen and Stoeva 2003] O. Christensen and D. T. Stoeva, "p-frames in separable Banach spaces", *Adv. Comput. Math.* **18**:2-4 (2003), 117–126. MR 2004b:42060 Zbl 1012.42024

[Daubechies et al. 1986] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions", *J. Math. Phys.* **27**:5 (1986), 1271–1283. MR 87e:81089 Zbl 0608.46014

[Douglas 1972] R. G. Douglas, *Banach algebra techniques in operator theory*, Pure and Applied Mathematics **49**, Academic Press, New York, 1972. MR 50 #14335 Zbl 0247.47001

[Dragomir 2004] S. S. Dragomir, *Semi-inner products and applications*, Nova Science, Hauppauge, NY, 2004. MR 2005b:46053 Zbl 1060.46001

[Duffin and Schaeffer 1952] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", *Trans. Amer. Math. Soc.* **72** (1952), 341–366. MR 13,839a Zbl 0049.32401

[Dunford and Schwartz 1958] N. Dunford and J. T. Schwartz, *Linear operators, I: General theory*, Pure and Applied Math. **7**, Interscience, New York, 1958. MR 22 #8302

[Fabian et al. 2001] M. Fabian, P. Habala, P. Hájek, V. Montesinos Santalucía, J. Pelant, and V. Zizler, *Functional analysis and infinite-dimensional geometry*, CMS Books in Mathematics **8**, Springer, New York, 2001. MR 2002f:46001 Zbl 0981.46001

[Gabardo and Han 2003] J.-P. Gabardo and D. Han, "Frames associated with measurable spaces", *Adv. Comput. Math.* **18**:2-4 (2003), 127–147. MR 2004b:42062 Zbl 1033.42036

[Han and Larson 2000] D. Han and D. R. Larson, *Frames, bases and group representations*, Mem. Amer. Math. Soc. **697**, American Mathematical Society, Providence, RI, 2000. MR 2001a:47013 Zbl 0971.42023

[Heil and Walnut 1989] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms", *SIAM Rev.* **31**:4 (1989), 628–666. MR 91c:42032 Zbl 0683.42031

[Heuser 1982] H. G. Heuser, *Functional analysis*, Wiley, New York, 1982. MR 83m:46001 Zbl 0465.47001

[Joveini and Amini 2009] R. Joveini and M. Amini, "Yet another generalization of frames and Riesz bases", *Involve* **2**:4 (2009), 395–407. MR 2010k:42060 Zbl 1184.42026

[Martin 1976] R. H. Martin, Jr., *Nonlinear operators and differential equations in Banach spaces*, Wiley, New York, 1976. MR 58 #11753 Zbl 0333.47023

[Najati and Faroughi 2007] A. Najati and M. H. Faroughi, "p-frames of subspaces in separable Hilbert spaces", *Southeast Asian Bull. Math.* **31**:4 (2007), 713–726. MR 2009d:46045 Zbl 1150.46011

[Rahimi et al. 2006] A. Rahimi, A. Najati, and Y. N. Dehghan, "Continuous frames in Hilbert spaces", *Methods Funct. Anal. Topology* **12**:2 (2006), 170–182. MR 2007d:42061

[Rudin 1973] W. Rudin, *Functional analysis*, McGraw-Hill, New York, 1973. MR 51 #1315 Zbl 0253.46001

[Rudin 1974] W. Rudin, *Real and complex analysis*, 2nd ed., McGraw-Hill, New York, 1974. MR 49 #8783 Zbl 0278.26001

[Stampfli 1969] J. G. Stampfli, "Adjoint abelian operators on Banach space", *Canad. J. Math.* **21** (1969), 505–512. MR 39 #807 Zbl 0183.14001

[Stoeva 2008] D. T. Stoeva, "Generalization of the frame operator and the canonical dual frame to Banach spaces", *Asian-Eur. J. Math.* **1**:4 (2008), 631–643. MR 2009m:42058

[Sun 2006] W. Sun, "G-frames and g-Riesz bases", *J. Math. Anal. Appl.* **322**:1 (2006), 437–452. MR 2007b:42047 Zbl 1129.42017

Received: 2011-02-17 Accepted: 2011-02-26

mhfaroughi@yahoo.com *Faculty of Mathematical Science, University of Tabriz, 29 Bahman Boulevard, Tabriz, Iran*

Department of Mathematics, Islamic Azad University, Shabestar Branch, Shabestar 0098, Iran

osgooei@tabrizu.ac.ir *Faculty of Mathematical Science, University of Tabriz, 29 Bahman Boulevard, Tabriz 0098, Iran*

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.187

The multidimensional Frobenius problem

Jeffrey Amos, Iuliana Pascu, Vadim Ponomarenko,
Enrique Treviño and Yan Zhang

(Communicated by Scott Chapman)

We provide a variety of results concerning the problem of determining maximal vectors g such that the Diophantine system $Mx = g$ has no solution: conditions for the existence of g, conditions for the uniqueness of g, bounds on g, determining g explicitly in several important special cases, constructions for g, and a reduction for M.

1. Introduction

Let m, x be column vectors from the nonnegative integers $\mathbb{N}_0$. Georg Frobenius focused attention on determining the maximal integer g such that the linear Diophantine equation $m^T x = g$ has no solutions. This problem has attracted substantial attention in the last 100 years; for a survey see [Ramírez Alfonsín 2006]. In this paper, we consider the problem of determining maximal vectors g such that the system of linear Diophantine equations $Mx = g$ has no solutions.

For any real matrix X and any $S \subseteq \mathbb{R}$, we write X_S for $\{Xs : s \in S^k\}$, where k denotes the number of columns of X. We write X_1 for the vector in $X_{\{1\}}$. We fix $M \in \mathbb{Z}_{n\times(n+m)}$, and write $M = [A|B]$, where A is $n \times n$. We call $A_{\mathbb{R}^{\geq 0}}$ the *cone*, and $M_{\mathbb{N}_0}$ the *monoid*. $|A|$ denotes henceforth the absolute value of $\det A$, if A is a square matrix; but still the cardinality of A, if A is a set. If $|A| \neq 0$, then we follow [Novikov 1992] and call the cone *volume*. If each column of B lies in the volume cone, then we call M *simplicial*. Unless otherwise noted, we assume henceforth that M is simplicial. Note that if $n \leq 2$ and there is some half-space containing all the columns of M, then we may always rearrange columns to make M simplicial. For $x \in \mathbb{R}^n$, we call $x + M_{\mathbb{R}^{\geq 0}} = x + A_{\mathbb{R}^{\geq 0}}$ the cone at x, writing $\mathrm{cone}(x)$.

Let $u, v \in \mathbb{R}^n$. If $u - v \in A_{\mathbb{Z}}$, then we write $u \equiv v$ and say that u, v are *equivalent mod* A. If $u - v \in A_{\mathbb{R}^{\geq 0}}$, then we write $u \geq v$. If $u - v \in A_{\mathbb{R}^{>0}}$, then we write $u \succ v$. Note that $u \succ v$ implies $u \geq v$, and $u \succ v \geq w$ implies $u \succ w$; however, $u \gtrneq v$ does

MSC2010: 11B75, 11D04, 11D72.
Keywords: Frobenius, coin-exchange, linear Diophantine system.
Research supported in part by NSF grant 0097366.

not imply that $u \succ v$. For $v \in \mathbb{R}^n$, we write $(v)_i$ for the i-th coordinate of v, and $[\succ v] = \{u \in \mathbb{Z}^n : u \succ v\}$. We say that v is *complete* if $[\succ v] \subseteq M_{\mathbb{N}_0}$. We set G, more precisely $G(M)$, to be the set of all $\geq$-minimal complete vectors. We call elements of G *Frobenius vectors*; they are the vector analogue of g that we will investigate.

Set $Q = (1/|A|)\mathbb{Z} \subseteq \mathbb{Q}$. Although G is defined in $\mathbb{R}^n$, in fact it is a subset of Q^n, by the following result. Furthermore, the columns of B are in $A_{Q^{\geq 0}}$; hence $M_{Q^{\geq 0}} = A_{Q^{\geq 0}}$ and without loss we work over Q rather than over $\mathbb{R}$.

Proposition 1.1. *Let $v \in \mathbb{R}^n$. There exists $v^\star \in Q^n$ with $[\succ v] = [\succ Av^\star]$ and $v \geq Av^\star$.*

Proof. We choose $v^\star \in Q^n$ such that $A^{-1}v - v^\star = \epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ with $0 \leq \epsilon_i < 1/|A|$. Multiplying by A we get $v - Av^\star = A\epsilon$; hence $v \geq Av^\star$. We will now show that for $u \in \mathbb{Z}^n$, $u \succ v$ if and only if $u \succ Av^\star$. If $u \succ v$, then $u \succ Av^\star$ because $u \succ v \geq Av^\star$. On the other hand, suppose that $u \succ Av^\star$ and $u \not\succ v$. Hence $u - Av^\star \in A_{\mathbb{R}^{>0}}$ and $u - v \in A_{\mathbb{R}} \setminus A_{\mathbb{R}^{>0}}$. Multiplying by A^{-1} we get $A^{-1}u - v^\star \in I_{\mathbb{R}^{>0}}$ and $A^{-1}u - A^{-1}v \in I_{\mathbb{R}} \setminus I_{\mathbb{R}^{>0}}$. Therefore, there is some coordinate i with $(A^{-1}u - v^\star)_i > 0$ and $(A^{-1}u - A^{-1}v)_i \leq 0$. Because $u \in \mathbb{Z}^n$ and A is an integer matrix, we have $A^{-1}u \in Q^n$; hence in fact $(A^{-1}u - v^\star)_i \geq 1/|A|$. Now, $0 \geq (A^{-1}u - A^{-1}v)_i = (A^{-1}u - v^\star - (A^{-1}v - v^\star))_i = (A^{-1}u - v^\star)_i - \epsilon_i \geq 1/|A| - \epsilon_i$. However, this contradicts $\epsilon_i < 1/|A|$. □

Let $x, y \in M_{Q^{\geq 0}}$. We write $x = Ax'$, $y = Ay'$, with $x', y' \in (Q^{\geq 0})^n$, define z' via $(z')_i = \max((x')_i, (y')_i)$, and set $\operatorname{lub}(x, y) = Az'$. We have $\operatorname{lub}(x, y) \in M_{Q^{\geq 0}}$, although in general $\operatorname{lub}(x, y) \notin M_{\mathbb{N}_0}$ (even if $x, y \in M_{\mathbb{N}_0}$) because $A^{-1}B$ need not have integer entries.

For $u \in M_Q$, we set $V(u) = (u + A_{Q \cap (0,1]}) \cap \mathbb{Z}^n$. It was known to Dedekind [1877] that $|V(u)| = |A|$, and that $V(u)$ is a complete set of coset representatives mod A (as restricted to $\mathbb{Z}^n$). Note that u is complete if and only if $V(u) \subseteq M_{\mathbb{N}_0}$.

The following equivalent conditions on M generalize the one-dimensional notion of relatively prime generators. Portions of the following have been repeatedly rediscovered [Frumkin 1981; Ivanov and Shevchenko 1975; Novikov 1992; Rycerz 2000; Vizvári 1987]. We assume henceforth, unless otherwise noted, that M possesses these properties. We call such M *dense*.

Theorem 1.2. *The following are equivalent*:

(1) *G is nonempty.*

(2) *$M_{\mathbb{Z}} = \mathbb{Z}^n$.*

(3) *For all unit vectors e_i $(1 \leq i \leq n)$, $e_i \in M_{\mathbb{Z}}$.*

(4) *There is some $v \in M_{\mathbb{N}_0}$ with $v + e_i \in M_{\mathbb{N}_0}$ for all unit vectors e_i.*

(5) *The GCD of all the $n \times n$ minors of M has absolute value* 1.

(6) *The elementary divisors of M are all* 1.

Proof. The proof follows the plan (1) $\iff$ (4) $\iff$ (3) $\iff$ (2) $\iff$ (6) $\iff$ (5).

(1) $\iff$ (4): Let $g \in G$. Choose $v \in [\succ g]$ far enough from the boundaries of the cone so that $v + e_i$ is also in $[\succ g]$ for all unit vectors e_i. Because g is complete, v and $v + e_i$ are all in $M_{\mathbb{N}_0}$. The other direction is proved in [Novikov 1992, Proposition 5].

(4) $\iff$ (3): For one direction, write $e_i = Mf_i$. Set $k = \max_i \| f_i \|_\infty$. Set $v = Mk^n$. We see that $v + e_i = M(k^n + f_i) \subseteq M_{\mathbb{N}_0}$. For the other direction, let $1 \le i \le n$. Write $v = Mw$, $v + e_i = Mw'$, where $w, w' \in \mathbb{N}_0^n$. Hence, $e_i = M(w' - w) \subseteq M_{\mathbb{Z}}$.

(3) $\iff$ (2): Let $v \in \mathbb{Z}^n$; write $v = (v_1, v_2, \ldots, v_n)$. Write $e_i = Mf_i$, for $f_i \in \mathbb{Z}^n$. Then $v = M \sum v_i f_i$, as desired. The other direction is trivial.

(2) $\iff$ (6): We place M in Smith normal form: write $M = LNR$, where N is a diagonal matrix of the same dimensions as M, and L, R are square matrices, invertible over the integers. The diagonal entries of N are the elementary divisors of M. We therefore have that (2) $\iff N = [I|0] \iff$ (6).

(6) $\iff$ (5): The product of the elementary divisors is known (see, for example, [van der Waerden 1967, Remark 3 in Section 12.2]) to be the absolute value of the GCD of all $n \times n$ minors of M. If they are each one, then their product is one. Conversely, if their product is one, then they must each be one since they are all nonnegative integers. □

Classically, there is a second type of Frobenius number f, maximal so that $m^T x = f$ has no solutions with x from $\mathbb{N}$ (rather than $\mathbb{N}_0$). This does not alter the situation; in [Brauer and Shockley 1962] it was shown that $f = g + m^T 1$. A similar situation holds in the vector context.

Call v *f-complete* if $[\succ v] \subseteq M_{\mathbb{N}}$.

Proposition 1.3. *Let F be the set of all $\geq$-minimal f-complete vectors. Then $F = G + M_1$.*

Proof. It suffices to show that $v \in Q^n$ is complete if and only if $v + M_1$ is f-complete. The following conditions are equivalent for an integral vector u: (1) $u \in [\succ v + M_1]$; (2) $u \succ v + M_1$; (3) $(u - M_1) - v \in M_{\mathbb{R}^{\geq 0}}$; (4) $(u - M_1) \succ v$; (5) $(u - M_1) \in [\succ v]$. Now, suppose that v is complete. Let $u \in [\succ v + M_1]$; hence $(u - M_1) \in [\succ v] \subseteq M_{\mathbb{N}_0}$ and therefore $u \in M_{\mathbb{N}}$. So $v + M_1$ is f-complete. On the other hand, suppose that $v + M_1$ is f-complete. Let $(u - M_1) \in [\succ v]$; hence $u \in [\succ v + M_1] \subseteq M_{\mathbb{N}}$. Hence $u - M_1 \subseteq M_{\mathbb{N}} - M_1 = M_{\mathbb{N}_0}$, and v is complete. □

Having established the notation and basic groundwork for the problem, we now present two useful techniques: the method of critical elements, and the MIN method. Each will be shown to characterize the set G.

2. The method of critical elements

For a vector u and $i \in [1, n]$, let

$$C^i(u) = \{v : v \in \mathbb{Z}^n \setminus M_{\mathbb{N}_0}, v = u + Aw, (w)_i = 0, (w)_j \in (0, 1] \text{ for } j \neq i\}.$$

This set captures all lattice points missing from the monoid, in the i-th face of the cone at u, that are minimal mod A. Let $C(u) = \bigcup_{i\in[1,n]} C^i(u)$, which is a disjoint union of finite sets. We call elements of $C(u)$ *critical*. Note that if $v \in C^i(u)$, then $v + Ae_i \in V(u)$. Critical elements characterize G, as shown by the following theorem.

Theorem 2.1. *Let x be complete. The following statemements are equivalent.*

(1) $x \in G$.

(2) *Each face of* $\operatorname{cone}(x)$ *contains at least one lattice point not in the monoid.*

(3) $C^i(x) \neq \varnothing$ *for all* $i \in [1, n]$.

Proof. We write $x = Ax'$. For each $i \in [1, n]$, set $x^i = x - (1/|A|)Ae_i$ and $S_i = [\succ x^i] \setminus [\succ x]$. Observe that $S_i = \{Au \in \mathbb{Z}^n : (u)_j > (x')_j \text{ (for } j \neq i), (u)_i = (x')_i\}$; the S_i are the lattice points in the i-th face of $\operatorname{cone}(x)$.

(1) $\Longrightarrow$ (2) If $S_i \subseteq M_{\mathbb{N}_0}$, then x^i is complete, which violates $x \in G$.

(2) $\Longrightarrow$ (3) Pick any minimal $y \in S_i \setminus M_{\mathbb{N}_0}$. Suppose that $(A^{-1}(y - x))_j \notin (0, 1]$ for $j \neq i$; in this case, $y - Ae_j$ would also be in $S_i \setminus M_{\mathbb{N}_0}$, violating the minimality of y. Hence $y \in C^i(x)$, and thus $C^i(x) \neq \varnothing$.

(3) $\Longrightarrow$ (1) If $x^\star < x$, then $x^\star \leq x^i$ for some i. But no x^i is complete; hence $x^\star$ is not complete. Thus x is $\geq$-minimal and complete and thus $x \in G$. □

Critical elements can also be used to test for uniqueness of Frobenius vectors. Set $\bar{e}_i = \overline{1} - e_i = (1, 1, \ldots, 1, 0, 1, 1, \ldots, 1)$.

Theorem 2.2. *Let $g \in G$. Then $|G| = 1$ if and only if for each $i \in [1, n]$ there is some $c^i \in C^i(g)$ with $c^i + kA\bar{e}_i \notin M_{\mathbb{N}_0}$ for all $k \in \mathbb{N}_0$.*

Proof. Suppose that for each $i \in [1, n]$ there is some $c^i \in C^i(g)$ with $c^i + kA\bar{e}_i \notin M_{\mathbb{N}_0}$ for all k. Let $g' \in G$. If $g' \neq g$, then for some i we must have $(A^{-1}g')_i < (A^{-1}g)_i$. As $k \to \infty$, $(A^{-1}c^i + k\bar{e}_i)_j \to \infty$ (for $j \neq i$), but also $(A^{-1}c^i + k\bar{e}_i)_i = (A^{-1}g)_i$ for all k. Therefore, for some k we have $c^i + kA\bar{e}_i \succ g'$. Hence g' is not complete, which is violative of our assumption. Hence $|G| = 1$.

Now, let $g \in G$ be unique, let $i \in [1, n]$ be such that each $c^i \in C^i(g)$ has some $k(i)$ with $c^i + k(i)A\bar{e}_i \in M_{\mathbb{N}_0}$. If $c^i + kA\bar{e}_i \in M_{\mathbb{N}_0}$, then $c^i + k'A\bar{e}_i \in M_{\mathbb{N}_0}$ for any $k' \geq k$; hence because $|C^i(g)| < \infty$ there is some $K \in \mathbb{N}_0$ with $c^i + KA\bar{e}_i \in M_{\mathbb{N}_0}$

for all $c^i \in C^i(g)$. Now, set

$$g^\star = g + (K+1)A\bar{e}_i - (1/|A|)Ae_i,$$
$$S = [\succ g^\star]\backslash[\succ g] \subseteq \{u \in \mathbb{Z}^n : (A^{-1}(u-g))_i = 0, (A^{-1}(u-g))_j \geq K+1 \ (j \neq i)\}.$$

We now show that $S\backslash M_{\mathbb{N}_0}$ is empty; otherwise, choose u therein. Set $u' = u - Aa$, where $(a)_i = 0$ and, for $j \neq i$,

$$(a)_j = \begin{cases} \lfloor (A^{-1}(u-g))_j \rfloor & \text{if } (A^{-1}(u-g))_j \notin \mathbb{Z}, \\ (A^{-1}(u-g))_j - 1 & \text{if } (A^{-1}(u-g))_j \in \mathbb{Z}, \end{cases}$$

Then $u' \in \mathbb{Z}^n \setminus M_{\mathbb{N}_0}$, since otherwise $u \in M_{\mathbb{N}_0}$. We also have $(A^{-1}(u'-g))_i = 0$ and $(A^{-1}(u'-g))_j \in (0, 1]$ for $j \neq i$; hence $u' \in C^i(g)$. But then $u' + KA\bar{e}_i \in M_{\mathbb{N}_0}$ and hence $u \in M_{\mathbb{N}_0}$ since $u - (u' + KA\bar{e}_i) \in A_{\mathbb{N}_0}$. Hence $S \subseteq M_{\mathbb{N}_0}$ and $g^\star$ is complete. Now take $g' \in G$ with $g' \leq g^\star$. We have $(A^{-1}g')_i \leq (A^{-1}g^\star)_i < (A^{-1}g)_i$ and hence $g' \neq g$, which is violative of our hypothesis. □

Our next result generalizes a one-dimensional reduction result in [Johnson 1960] which is very important because it allows the assumption that the generators are pairwise relatively prime. The vector generalization unfortunately does not permit us an analogous assumption in general.

Theorem 2.3. *Let $d \in \mathbb{N}$ and let $M = [A|B]$ be simplicial. Suppose that $N = [A|dB]$ is dense. Then M is dense, and $G(N) = dG(M) + (d-1)A_1$.*

Proof. Each $n \times n$ minor of M divides a corresponding minor of N, and hence M is dense. Further, d divides all minors of N apart from $|A|$, and hence $\gcd(|A|, d) = 1 = \gcd(|A|^2, d)$. We can therefore pick $d^\star \in \mathbb{N}$ with $d^\star d \in 1 + |A|^2\mathbb{N}_0$. For any $v \in Q^n$, we observe that $d^\star dv - v \in \mathbb{N}_0|A|^2Q^n = \mathbb{N}_0|A|\mathbb{Z}^n \subseteq A_{\mathbb{Z}}$; hence $d^\star dv \equiv v$. Set $\theta(x) = dx + (d-1)A1^n$. We will show for any $x \in Q^n$ that $x \in M_{\mathbb{N}_0}$ if and only if $\theta(x) \in N_{\mathbb{N}_0}$ (in particular, if $\theta(x) \in N_{\mathbb{N}_0}$, then $x \in \mathbb{Z}^n$). One direction is trivial; for the other, assume $\theta(x) \in N_{\mathbb{N}_0}$. We have $dx + dA1^n = A(y + 1^n) + dBz$, for $y \in \mathbb{N}_0^n$, and $z \in \mathbb{N}_0^m$. We observe that $x + A1^n = A(1/d)(y + 1^n) + Bz$, so $x + A1^n \geq Bz$. Also, $d^\star d(x + A1^n) = Ad^\star(y + 1^n) + d^\star dBz$, and hence $x + A1^n \equiv Bz$. Therefore $x + A1^n - Bz = Aw$ for some $w \in \mathbb{N}_0^n$. Further, $w = (1/d)(y + 1^n)$ so in fact $w \in \mathbb{N}^n$. Hence, $x = A(w - 1^n) + Bz \in M_{\mathbb{N}_0}$.

Next, we show that x is M-complete if and only if $\theta(x)$ is N-complete. First suppose that $\theta(x)$ is N-complete. Let $u \in [\succ x]$; we have $\theta(u) \in [\succ \theta(x)] \subseteq N_{\mathbb{N}_0}$. Hence $u \in M_{\mathbb{N}_0}$ so x is M-complete. Now suppose that x is M-complete. Let $u \in V(\theta(x))$. Set $u' \in V(x)$ with $du' \equiv u$. We have $u = \theta(x) + A\epsilon$, $u' = x + A\epsilon'$, where $\epsilon, \epsilon' \in (0, 1]^n$. We compute $u - du' = A\omega$, where $\omega = d(1^n - \epsilon') + (\epsilon - 1^n)$. Because $u \equiv du'$ we also have $u - du' = A\alpha$ with $\alpha \in \mathbb{Z}^n$. Since $|A| \neq 0$, we have $\omega = \alpha \in \mathbb{Z}^n$. Further, since $\epsilon, \epsilon' \in (0, 1]^n$, each coordinate of $d(1^n - \epsilon') + (\epsilon - 1^n)$

is strictly greater than -1 and hence $\omega \in \mathbb{N}_0^n$. We have $u' \in M_{\mathbb{N}_0}$ since x is M-complete. But then $du' \in N_{\mathbb{N}_0}$, and thus $u = du' + A\omega \in N_{\mathbb{N}_0}$. Hence $V(\theta(x)) \subseteq N_{\mathbb{N}_0}$ and thus $\theta(x)$ is N-complete.

Let $g \in G(M)$. We will show that $\theta(g) \in G(N)$. Let $i \in [1, n]$. By Theorem 2.1, there is $u \in [0, 1]^n$ with $u_i = 0$, $u_j > 0$ (for $j \neq i$), such that $g + Au \in \mathbb{Z}^n \setminus M_{\mathbb{N}_0}$. We have $\theta(g + Au) \in \mathbb{Z}^n \setminus N_{\mathbb{N}_0}$. We write $\theta(g + Au) = d(g + Au) + (d-1)A1^n = \theta(g) + Adu$. Write $du = u' + u''$ where $(u')_i = 0$, $(u')_j \in (0, 1]$, and $u'' \in \mathbb{N}_0^n$. We have $\theta(g) + Au' \in C^i(\theta(g))$; considering all i gives $\theta(g) \in G(N)$. Now, let $g \in G(N)$. We will show that $\theta^{-1}(g) = (1/d)(g - (d-1)A1^n) \in G(M)$. We again apply Theorem 2.1 to get an appropriate u with $g + Au \in \mathbb{Z}^n \setminus N_{\mathbb{N}_0}$. Note that $g + A(u + d1^n) \in N_{\mathbb{N}_0}$; hence

$$\begin{aligned}\theta^{-1}(g + A(u + d1^n)) &= (1/d)(g + Au + dA1^n - (d-1)A1^n) \\ &= \theta^{-1}(g) + (1/d)Au + A1^n \in M_{\mathbb{N}_0} \subseteq \mathbb{Z}^n.\end{aligned}$$

Thus, $\theta^{-1}(g + Au) = (1/d)(g + Au - (d-1)A1^n) = \theta^{-1}(g) + (1/d)Au \in \mathbb{Z}^n$. We therefore have $\theta^{-1}(g + Au) \in C^i(\theta^{-1}(g))$; considering all i gives $\theta^{-1}(g) \in G(M)$. □

3. The MIN method

Let $\text{MIN} = \{x : x \in M_{\mathbb{N}_0};\ \text{for all } y \in M_{\mathbb{N}_0},\ \text{if } y \equiv x \text{ then } y \geq x\}$. Provided M is dense, MIN will have at least one representative of each of the $|A|$ equivalence classes mod A. MIN is a generalization of a one-dimensional method in [Brauer and Shockley 1962]; the following result shows that it characterizes the set G.

Theorem 3.1. *Let $g \in G$. Then $g = \text{lub}(N) - A_1$ for some complete set of coset representatives $N \subseteq \text{MIN}$. Further, if $n < |A|$ then there is some $N' \subseteq N$ with $|N'| = n$ and $\text{lub}(N) = \text{lub}(N')$.*

Proof. Observe that $V(g) \subseteq [\succ g]$, and hence $V(g) \subseteq M_{\mathbb{N}_0}$ since g is complete. Let $\text{MIN}' = \{u \in \text{MIN} : \exists v \in V(g), u \equiv v, u \leq v\}$. Now, for $v \in C^i(g)$, we have $v + Ae_i \in V(g)$. Let $v_{\text{MIN}} \in \text{MIN}'$ with $v_{\text{MIN}} \equiv v + Ae_i$ and $v_{\text{MIN}} \leq v + Ae_i$. We must have $(A^{-1}v_{\text{MIN}})_i \geq (A^{-1}v)_i + 1 = (A^{-1}g)_i + 1$ because otherwise $v \in v_{\text{MIN}} + A_{\mathbb{N}_0}$ and therefore $v \in M_{\mathbb{N}_0}$, which is violative of $v \in C^i(g)$. Set $N' = \{v_{\text{MIN}} : i \in [1, n]\}$. We have $\text{lub}(N') \geq g + A_1$, but also we have $g + A_1 = \text{lub}(V(g)) \geq \text{lub}(\text{MIN}') \geq \text{lub}(N')$. Hence all the inequalities are equalities, and in fact $\text{lub}(N') = \text{lub}(N)$ for any N with $N' \subseteq N \subseteq \text{MIN}'$. Finally, we note that $|N'| \leq n$ but also we may insist that $|N'| \leq |A|$ because $|V(g)| = |A|$. □

Elements of MIN have a particularly nice form. This is quite useful in computations.

Theorem 3.2. $\text{MIN} \subseteq \{Bx : x \in \mathbb{N}_0^m, \|x\|_1 \leq |A| - 1\}$.

Proof. Let $v \in \mathrm{MIN} \subseteq M_{\mathbb{N}_0}$. Write $v = Mv'$, where $v' \in \mathbb{N}_0^{n+m}$. Suppose that $(v')_i > 0$, for $1 \le i \le n$. Set $w' = v' - e_i$, and $w = Mw'$. We see that $w \equiv v$, $w \le v$, and $w \in M_{\mathbb{N}_0}$; this contradicts that $v \in \mathrm{MIN}$. Hence $\mathrm{MIN} \subseteq B_{\mathbb{N}_0}$. Let $z = Bx \in \mathrm{MIN}$. Suppose that $\|x\|_1 \ge |A|$. Start with 0 and increment one coordinate at a time, building a sequence $B0 = Bv_0 \lneq Bv_1 \lneq Bv_2 \lneq \cdots \lneq Bv_{\|x\|_1} = z$ where each $v_i \in \mathbb{N}_0^m$. We may do this since M is simplicial. Because there are at least $|A|+1$ terms, two (say $Bv_a \lneq Bv_b$) are congruent mod A. We have $z - Bv_b \in M_{\mathbb{N}_0}$ and so $y = z - (Bv_b - Bv_a) \in M_{\mathbb{N}_0}$, but $y \lneq z$ and $y \equiv z$. This violates that $z \in \mathrm{MIN}$. □

Corollary 3.3. *$|G|$ is finite.*

The following result, proved first in [Knight 1980] and rediscovered in [Simpson and Tijdeman 2003], generalizes the classical one-dimensional result on two generators that $g(a_1, a_2) = a_1a_2 - a_1 - a_2$. Note that in the special case where $m = 1$, we must have that $|G| = 1$ and $G \subseteq \mathbb{Z}^n$. Neither of these necessarily holds for $m > 1$.

Corollary 3.4. *If $m = 1$ then $G = \{|A|B - A_1 - B\}$.*

Proof. By Theorem 3.2, we have $\mathrm{MIN} = \{0, B, 2B, \ldots, (|A|-1)B\}$, a complete set of coset representatives. By Theorem 3.1, any $g \in G$ must have $g + A_1 = \mathrm{lub}(MIN) = (|A|-1)B$. □

Corollary 3.4 can be extended to the case where the column space of B is one dimensional, using as an oracle function the (one-dimensional) Frobenius number. In this special case we again have $|G| = 1$ and $G \subseteq \mathbb{Z}^n$.

Theorem 3.5. *Consider a dense $M = [A|B]$ with B a column $(n \times 1)$ vector, i.e., the special case $m = 1$. Let $C = [c_1, c_2, \ldots, c_m] \in \mathbb{N}^m$. Suppose that $P = [\,|A|\,|\,C\,]$ is dense. Then $N = [A|BC]$ is dense, and $G(N) = \{G(P)B + |A|B - A_1\}$.*

Proof. By Theorem 3.2, we have $\mathrm{MIN}(M) = \{0, B, \ldots, (|A|-1)B\}$. Hence $\mathbb{Z}^n/A\mathbb{Z}^n$ is cyclic, and B is a generator. Let S denote the set of all $n \times n$ minors of M, apart from $|A|$. Using the denseness of M and P, we have

$$\begin{aligned}\gcd(|A|, \{c_i s : 1 \le i \le m, s \in S\}) &= \gcd(|A|, \gcd(c_1, c_2, \ldots, c_m)\gcd(S)) \\ &= \gcd(|A|, \gcd(S)) = 1;\end{aligned}$$

hence N is dense. Again by Theorem 3.2, we have $\mathrm{MIN}(N) \subseteq B_{\mathbb{N}_0}$. We now show that $G(P)B \notin M_{\mathbb{N}_0}$. Suppose otherwise. We then write $G(P)B = Ax + BCy$ and hence $Ax = Bq$ for $q = (G(P) - Cy)$. We conclude that $qB \equiv 0 \bmod A$ and hence $q = k|A|$ for some $k \in \mathbb{N}$ ($k > 0$ since M is simplicial) since B generates $\mathbb{Z}^n/A\mathbb{Z}^n$. We now have $BG(P) = Bk|A| + BCy$, and hence $G(P) = k|A| + Cy$. But now $G(P) - 1$ is complete (with respect to P), which violates the definition of $G(P)$. Therefore $G(P)B \notin M_{\mathbb{N}_0}$. On the other hand, if $\alpha \in \mathbb{Z}$ and $\alpha > G(P)$ we have $\alpha = k|A| + Cy$, for some $k, y \in \mathbb{N}_0$. Therefore, we have $B\alpha = k|A|B + BCy =$

$A\left(k|A|A^{-1}B\right)+BCy \in M_{\mathbb{N}_0}$ (note that $A^{-1}B \in Q^{\geq 0}$ since M is simplicial). Hence, $T = \{G(P)B + kB : k \in [1, |A|]\} \subseteq M_{\mathbb{N}_0}$, with $\text{lub}(T) = G(P)B + |A|B = \beta$. Let $g \in G(N)$, and let M be chosen as in Theorem 3.1 with $|M| = |A|$. Since T is a complete set of coset representatives and both T and $\text{MIN}(N)$ lie on $B\mathbb{R}$, we have $\text{lub}(M) \leq \text{lub}(\text{MIN}(N)) \leq \text{lub}(T) = G(P)B + |A|B = \beta$. However, the coset of β is precisely $\{G(P)B + k|A|B : k \in \mathbb{Z}\}$. Therefore, β is the unique representative of its equivalence class in MIN, and thus $\beta \in M$ and $\text{lub}(M) = \beta$. Hence $g + A_1 = \beta$ for all $g \in G$, as desired. □

Example 3.6. Consider $N = \left(\begin{smallmatrix} 5 & 0 & 84 & 105 \\ 0 & 4 & 84 & 105 \end{smallmatrix}\right)$. We have $N = [A|BC]$, for $A = \left(\begin{smallmatrix} 5 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 3 \\ 3 \end{smallmatrix}\right)$, and $C = (28, 35)$. Following Theorem 3.5, we have $P = (20, 28, 35)$. $\gcd(20, 28, 35) = 1$ so P is dense; we now calculate $G(P) = 197$ using our one-dimensional oracle. Therefore N is dense and $G(N) = \left\{ \left(\begin{smallmatrix} 646 \\ 647 \end{smallmatrix}\right) \right\}$.

We give three more results using this method. First, we present a $\leq$-bound for G. This generalizes a one dimensional bound, attributed to Schur in [Brauer 1942]: $g(a_1, a_2, \ldots, a_k) \leq a_1a_k - a_1 - a_k$ (where $a_1 < a_2 < \cdots < a_k$). Note that Corollary 3.4 shows that equality is sometimes achieved.

Theorem 3.7. *For all $g \in G$, $g \leq \text{lub}\left(\{|A|b - A_1 - b : b \text{ a column of } B\}\right)$.*

Proof. Let $x \in \text{MIN}$, fix $1 \leq i \leq n$, and write

$$(A^{-1}x)_i = (A^{-1}Bx')_i = \left(\sum_b (x')_b A^{-1}b\right)_i,$$

where b ranges over all the columns of B. Set $b^\star$ to be a column of B with $(A^{-1}b^\star)_i$ maximal. By Theorem 3.2, we have that $(A^{-1}x)_i \leq (A^{-1}b^\star)_i \|x'\|_1 \leq (A^{-1}b^\star)_i(|A| - 1)$. By the choice of $b^\star$, and by varying i, we have shown that $x \leq \text{lub}(\{(|A| - 1)b\})$ and hence $\text{lub}(\text{MIN}) \leq \text{lub}(\{(|A| - 1)b\})$. For any $g \in G$, we apply Theorem 3.1 and have $g + A_1 \leq \text{lub}(\text{MIN}) \leq \text{lub}(\{(|A| - 1)b\})$. □

Next, we characterize possible G in our context for the special case $m = 1$. This generalizes a one-dimensional construction found in [Rosales et al. 2004]. If we allow $m = 2$, then it is an open problem to determine whether all G are possible.

Theorem 3.8. *Let $g \in \mathbb{Z}^n$. There exists a simplicial, dense, M with $m = 1$ and $G = \{g\}$ if and only if $\frac{1}{2}g \notin \mathbb{Z}^n$.*

Proof. Suppose $\frac{1}{2}g \notin \mathbb{Z}^n$. By applying an invertible change of basis, if necessary, we assume without loss that $g \in \mathbb{N}^n$ and that $\frac{1}{2}(g)_1 \notin \mathbb{Z}$. Set $A = \text{diag}(2, 1, 1, \ldots, 1)$, and set $B = A_1 + g$. For $i \in [1, n]$, define $A^{\underline{i}}$ to be A with the i-th column replaced by B. Note that $\det A = 2$ and $\det A^{\underline{1}} = 2 + (g)_1$ (which is odd), and hence M is dense. We now apply Corollary 3.4 to get $G = \{g\}$, as desired. Suppose now that we have a simplicial dense M, with $G = \{g\}$ and $\frac{1}{2}g \in \mathbb{Z}^n$. Applying Corollary 3.4 again, we get that $g + A_1 = (|A| - 1)B$. Suppose that $|A|$ were odd. Then each

coordinate of $(|A|-1)B$ is even, as is each coordinate of g, and hence so is each coordinate of A_1. Considering the integers mod 2, we have $|A| = 1$ but $A_1 = 0^n$, a contradiction. Therefore we must have that $|A|$ is even. We now consider the system $A(x_1, x_2, \ldots, x_n)^T = B$. We may apply Cramer's rule since $|A| \neq 0$ and $B \neq 0^n$; we find that, uniquely, $\det A^{\underline{i}} = x_i|A|$. We now consider the system reduced mod 2 (working in $\mathbb{Q}/2\mathbb{Q}$) and find that 1^n solves the reduced system, as $B = |A|B - g - A_1 \equiv -A1^n \equiv A1^n \pmod 2$. Hence, each x_i is in fact an odd integer, and thus $\det A^{\underline{i}}$ is an even integer. Consequently, all $n \times n$ minors of M are even, which is violative of the denseness of M. □

Our last result combines the two methods presented. It generalizes the one-dimensional theorem $g(a, a+c, a+2c, \ldots, a+kc) = a\lceil (a-1)/k \rceil + ac - a - c$, as proved in [Roberts 1956]. The following determines G, for M of a similarly special type.

Theorem 3.9. *Fix A and a vector $c \geq 0$. Set $C = c(1^n)^T$, a square matrix, and fix $k \in \mathbb{N}$. Set $M = [A|A+C|A+2C|\cdots|A+kC]$. Suppose that M is dense. Then $G(M) = \{Ax + |A|c - A_1 - c : x \in \mathbb{N}_0^n, \|x\|_1 = \lceil (|A|-1)/k \rceil\}$.*

Proof. We have

$$M_{\mathbb{N}_0} = \left\{ \sum_{i=0}^{k} (A + iC)x^i : x^i \in \mathbb{N}_0^n \right\} = \left\{ A \sum_{i=0}^{k} x^i + C \sum_{i=0}^{k} i x^i : x^i \in \mathbb{N}_0^n \right\}$$
$$= \left\{ A \sum_{i=0}^{k} x^i + c \sum_{i=0}^{k} i \|x^i\|_1 : x^i \in \mathbb{N}_0^n \right\}$$
$$= \left\{ Ax + c \sum_{i=0}^{k} i \|x^i\|_1 : x^i \in \mathbb{N}_0^n; x = \sum_{i=0}^{k} x^i \right\}.$$

Now, for a fixed $x \in \mathbb{N}_0^n$, as we vary the decomposition $x = \sum_{i=0}^{k} x^i$ (for $x^i \in \mathbb{N}_0^n$), we find that $\sum_{i=0}^{k} i\|x^i\|_1$ takes on all values from 0 to $k\|x\|_1$. Hence $M_{\mathbb{N}_0} = \{Ax + c\gamma : x \in \mathbb{N}_0^n, \gamma \in \mathbb{N}_0, \gamma \leq k\|x\|_1\}$.

Choose any $x \in \mathbb{N}_0^n$ satisfying $\|x\|_1 = \lceil (|A|-1)/k \rceil$. Set $T = \{Ax + c\gamma \in M_{\mathbb{N}_0} : 0 \leq \gamma \leq |A| - 1\}$. By construction, we have $T \subseteq M_{\mathbb{N}_0}$. Further, the elements of T must be inequivalent mod A, since c is a generator of the cyclic group $\mathbb{Z}^n/A_{\mathbb{Z}}$. Set $h = \operatorname{lub}(T) - A_1 = Ax + (|A|-1)c - A_1$. Note that each $t \in T$ either has $t \in V(h)$ or $t \leq t'$ (and $t \equiv t'$) for some $t' \in V(h)$; hence $V(h) \subseteq M_{\mathbb{N}_0}$ and h is complete. For any $i \in [1, n]$, $|A|-1 > k\|x - e_i\|_1$, so $A(x-e_i) + (|A|-1)c \in C^i(h)$, and thus $h \in G(M)$. Now, let $g \in G(M)$. By Theorem 3.1, we have $g \geq Ax + (|A|-1)c - A_1$, for some $x \in \mathbb{N}_0^n$ with $|A| - 1 \leq k\|x\|_1$. By our earlier observation, $Ax + (|A|-1)c - A_1 \in G(M)$, so we have equality by the minimality of g. □

Example 3.10. Consider $M = \left(\begin{smallmatrix} 5 & 0 & 7 & 2 & 9 & 4 & 11 & 6 & 13 & 8 & 15 & 10 & 17 & 12 & 19 & 14 \\ 0 & 4 & 1 & 5 & 2 & 6 & 3 & 7 & 4 & 8 & 5 & 9 & 6 & 10 & 7 & 11 \end{smallmatrix}\right)$. We see that $M = [A|A+C|A+2C|A+3C|A+4C|A+5C|A+6C|A+7C]$ for $A = \left(\begin{smallmatrix} 5 & 0 \\ 0 & 4 \end{smallmatrix}\right)$ and

$C = \left(\begin{smallmatrix} 2 & 2 \\ 1 & 1 \end{smallmatrix}\right)$. M is dense since $|A| = 20$, $|A+C| = 33$ and $\gcd(20, 33) = 1$. Applying Theorem 3.9, we get $G(M) = \left\{Ax + \binom{33}{15} : x,\ \|x\|_1 = 3\right\} = \left\{\binom{48}{15}, \binom{43}{19}, \binom{38}{23}, \binom{33}{27}\right\}$.

Acknowledgements

The authors would like to gratefully acknowledge the helpful comments of the anonymous referees.

References

[Brauer 1942] A. Brauer, "On a problem of partitions", *Amer. J. Math.* **64** (1942), 299–312. MR 3,270d Zbl 0061.06801

[Brauer and Shockley 1962] A. Brauer and J. E. Shockley, "On a problem of Frobenius", *J. Reine Angew. Math.* **211** (1962), 215–220. MR 26 #6113 Zbl 0108.04604

[Dedekind 1877] R. Dedekind, "Sur la théorie des nombres entiers algébriques", *Darboux Bull.* **9** (1877), 278. Translated as *Theory of algebraic integers*, Cambridge University Press, 1996.

[Frumkin 1981] M. A. Frumkin, "On the number of nonnegative integer solutions of a system of linear Diophantine equations", pp. 95–108 in *Studies on graphs and discrete programming* (Brussels, 1979), edited by P. Hansen, Ann. Discrete Math. **11**, North-Holland, Amsterdam, 1981. MR 83k:10028 Zbl 0477.90048

[Ivanov and Shevchenko 1975] N. N. Ivanov and V. N. Shevchenko, "The structure of a finitely generated semilattice", *Dokl. Akad. Nauk BSSR* **19**:9 (1975), 773–774. In Russian. MR 52 #8075 Zbl 0312.10009

[Johnson 1960] S. M. Johnson, "A linear diophantine problem", *Canad. J. Math.* **12** (1960), 390–398. MR 22 #12074 Zbl 0096.02803

[Knight 1980] M. J. Knight, "A generalization of a result of Sylvester's", *J. Number Theory* **12**:3 (1980), 364–366. MR 81j:10019 Zbl 0441.10010

[Novikov 1992] B. V. Novikov, "On the structure of subsets of a vector lattice that are closed with respect to addition", *Ukrain. Geom. Sb.* 35 (1992), 99–103. In Russian; translated in *J. Math. Sci.* **72**:4 (1994), 3223–3225. MR 95b:52027 Zbl 0850.06010

[Ramírez Alfonsín 2006] J. L. Ramírez Alfonsín, *The Diophantine Frobenius problem*, Oxford Lecture Series in Mathematics and its Applications **30**, Oxford University Press, Oxford, 2006. MR 2007i:11052 Zbl 1134.11012

[Roberts 1956] J. B. Roberts, "Note on linear forms", *Proc. Amer. Math. Soc.* **7** (1956), 465–469. MR 19,1038d Zbl 0071.03902

[Rosales et al. 2004] J. C. Rosales, P. A. García-Sánchez, and J. I. García-García, "Every positive integer is the Frobenius number of a numerical semigroup with three generators", *Math. Scand.* **94**:1 (2004), 5–12. MR 2004j:20117 Zbl 1077.20071

[Rycerz 2000] A. Rycerz, "The generalized residue classes and integral monoids with minimal sets", *Opuscula Math.* 20 (2000), 65–69. MR 2002k:11035

[Simpson and Tijdeman 2003] R. J. Simpson and R. Tijdeman, "Multi-dimensional versions of a theorem of Fine and Wilf and a formula of Sylvester", *Proc. Amer. Math. Soc.* **131**:6 (2003), 1661–1671. MR 2004k:11025 Zbl 1013.05087

[Vizvári 1987] B. Vizvári, "An application of Gomory cuts in number theory", *Period. Math. Hungar.* **18**:3 (1987), 213–228. MR 89d:11017 Zbl 0626.10013

[van der Waerden 1967] B. L. van der Waerden, *Algebra*, vol. 2, 5th ed., Heidelberger Taschenbücher **23**, Springer, Berlin, 1967. In German; translated as *Algebra*, vol. 2, Frederick Ungar, New York, 1970. MR 41 #8187b

Received: 2011-03-02 Revised: 2011-05-09 Accepted: 2011-05-09

jmamos1984@gmail.com *Department of Mathematics, Kansas State University, Manhattan, KS 66506, United States*

ipascu@wellesley.edu *Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142, United States*

vadim123@gmail.com *Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego CA 92182-7720, United States* http://www-rohan.sdsu.edu/~vadim/

enrique.trevino@dartmouth.edu *Department of Mathematics, Dartmouth College, Hanover, NH 03755, United States*

yanzhang@fas.harvard.edu *Department of Mathematics, Massachusetts Institute of Technology, 50 Memorial Drive, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, United States*

INVOLVE 4:2 (2011)
https://doi.org/10.2140/involve.2011.4.199

The Gauss–Bonnet formula on surfaces with densities

Ivan Corwin and Frank Morgan

(Communicated by Michael Dorff)

The celebrated Gauss–Bonnet formula has a nice generalization to surfaces with densities, in which both arclength and area are weighted by positive functions. Surfaces with densities, especially when arclength and area are weighted by the same factor, appear throughout mathematics, including probability theory and Perelman's recent proof of the Poincaré conjecture.

A classic, if somewhat anthropomorphic, question in mathematics is whether an ant moving on a curve embedded in $\mathbb{R}^3$ or in a surface can measure the curvature κ of the curve or say anything about how the curve is embedded in space. The answer, no, stems from the fact that the ant can only measure distance along the curve and has no way to determine changes in direction. Curvature is extrinsic to a curve and must be measured from outside the curve.

Following this one might then ask whether a person moving in a surface embedded in $\mathbb{R}^3$ has any chance of saying something about the surface's curvature in $\mathbb{R}^3$. Whereas the ant could only measure distance along the curve, a person on a surface has the ability to measure both length and area on the surface. Does this change things?

The answer is yes. Gauss's Theorem Egregium declares that a certain measure of surface curvature now known as the Gauss curvature G turns out to be an intrinsic quantity, measurable from within the surface. This is not at all apparent from its definition. G is defined as the product of the principal curvatures κ_1, κ_2, the largest and smallest (or most positive and most negative) curvatures of one-dimensional slices by planes orthogonal to the surface. For a plane, $G = 0$. For a sphere of radius a, we have $G = 1/a^2$. For the hyperbolic paraboloid $\{z = \frac{1}{2}(x^2 - y^2)\}$, at the origin G equals -1: negative because the surface is curving up in one direction and

MSC2010: 53B20.
Keywords: Gauss–Bonnet, density.
The authors acknowledge partial support by the National Science Foundation (research grant and graduate research fellowship).

down in the other direction; as you move farther out in the surface, G approaches 0 as the surface flattens out.

The fact that the Gauss curvature is actually intrinsic is a consequence of the celebrated Gauss–Bonnet formula (for a general reference see [do Carmo 1976; Morgan 1998]). Gauss–Bonnet relates the integral of the Gauss curvature over a smooth topological disc D in a surface to the integral over the boundary ∂D of the curvature κ of the boundary:

$$\int_{\partial D} \kappa + \int_D G = 2\pi.$$

For example, for a smooth closed curve C in the plane, where $G = 0$,

$$\int_C \kappa = 2\pi,$$

that is, the total curvature of an embedded planar curve is 2π. For a smooth closed curve C enclosing area A on the unit sphere, where $G = 1$,

$$\int_C \kappa + A = 2\pi.$$

For example, the equator, with curvature $\kappa = 0$, encloses area 2π. Note that we are using the intrinsic or "geodesic" curvature κ, not the curvature of the curve in $\mathbb{R}^3$ if the surface is embedded in $\mathbb{R}^3$.

Gauss–Bonnet has extensive applications throughout geometry and topology. It can be used to classify two-dimensional surfaces by genus and to solve isoperimetric problems [Howards et al. 1999; Morgan 1998, Section 9.12]. The Gauss–Bonnet formula provides an intrinsic definition of the Gauss curvature G of a surface at a point p by considering ϵ-balls B_ϵ of area A about p and taking a limit as ϵ approaches 0:

$$G(p) = \frac{1}{A}\int_{B_\epsilon} G = \lim \frac{1}{A}\Big(2\pi - \int_{\partial B_\epsilon} \kappa\Big).$$

This article considers what happens to the Gauss–Bonnet formula under some simple intrinsic alterations of the surface. The most common alteration, called a conformal change of metric, scales distance by a variable factor λ, so that $ds = \lambda\, ds_0$ and $dA = \lambda^2\, dA_0$; that is, arc length is weighted by λ and area is weighted by λ^2. More generally, one can weight arc length and area by unrelated densities:

$$ds = \delta_1 ds_0, \quad dA = \delta_2\, dA_0.$$

If the two densities are equal, $\delta_1 = \delta_2 = \Psi$, the result is simply called a surface with density Ψ. Surfaces with density appear throughout mathematics, including probability theory and Perelman's recent proof of the Poincaré conjecture [Morgan

2009, Chapter 18]. Important examples include quotients of Riemannian manifolds by symmetries and Gauss space, defined as $\mathbb{R}^n$ with Gaussian density $c\exp(-r^2)$.

Perelman's paper and many other applications require generalizations of curvature to general dimensional surfaces with densities. In higher dimensions, the important intrinsic curvature is the so-called Ricci curvature, for which many generalizations have been proposed, each for its own purpose, one particular choice employed by Perelman (see [Morgan 2009, Section 18.3] and references therein). Corwin et al. [2006, Section 5] proposed a generalization of Gauss curvature and the Gauss–Bonnet formula to surfaces with density Ψ. In principle, their definition generalizes to surfaces with length density δ_1 and area density δ_2 by a conformal change of metric. The following proposition gives a simple, direct presentation of that generalization. The generalized Gauss curvature G' is given by

$$G' = G - \Delta \log \delta_1.$$

An intriguing feature is that G' depends only on the length density δ_1, not on the area density δ_2. For a conformal change of metric ($\delta_1 = \lambda, \delta_2 = \lambda^2$), (1) below agrees with the standard Gauss–Bonnet formula (and gives an easy proof): the first integrand becomes $\kappa\lambda ds_0 = \kappa ds$ and the second integrand becomes the new Gauss curvature $G'\lambda^2 dA_0 = G' dA$ because $G' = (G - \Delta \log \lambda)/\lambda^2$ [Dubrovin et al. 1992, Theorem 13.1.3].

For a disc with density (the case $\delta_2 = \delta_1$), (1) agrees with the formula in [Corwin et al. 2006, Proposition 5.2]. For a disc with area density (the case $\delta_1 = 1$), (1) agrees with the formula in [Carroll et al. 2008, Proposition 3.3].

There are other possible generalizations of Gauss curvature to surfaces with density, for example, coming from the power series expansions for the area and perimeter of geodesic balls [Corwin et al. 2006, Propositions 5.8 and 5.9].

Proposition. *Consider a smooth Riemannian disc D with Gauss curvature G, length density δ_1, area density δ_2, classical boundary curvature κ_0 (inward normal), and hence generalized boundary curvature*

$$\kappa = (\delta_1/\delta_2)\kappa_0 - (1/\delta_2)\partial\delta_1/\partial n.$$

Then

$$\int_{\delta_D} (\delta_2/\delta_1)\kappa ds_0 + \int_D (G - \Delta \log \delta_1)\, dA_0 = 2\pi. \tag{1}$$

Proof. We begin by explaining the formula for κ. The geometric interpretation of curvature is minus the rate of change of length per change in enclosed area as you deform the curve normal to itself [Corwin et al. 2006, Proposition 3.2]. First of all, the densities weight this effect by δ_1/δ_2. There is a second effect due to the rate of change $\partial\delta_1/\partial n$ of the length density in the normal direction, divided again by the area density δ_2.

To prove (1), first consider the conformal metric $ds = \delta_1 ds_0$, with area density δ_1^2 and curvature

$$\kappa' = (1/\delta_1)\kappa_0 - (1/\delta_1^2)\partial\delta_1/\partial n.$$

Multiplying the area density by $\mu = \delta_2/\delta_1^2$ multiplies the curvature by $1/\mu = \delta_1^2/\delta_2$:

$$\kappa = (\delta_1/\delta_2)\kappa_0 - (1/\delta_2)\partial\delta_1/\partial n.$$

Hence by substitution, by the classical Gauss–Bonnet Theorem and the divergence theorem, and by trivial algebra,

$$\begin{aligned}\int_{\partial D} (\delta_2/\delta_1)\kappa\, ds_0 &= \int_{\partial D} \kappa_0 ds_0 - \int_{\partial D} \partial \log \delta_1/\partial n\, ds_0 \\ &= 2\pi - \int_D G\, dA_0 + \int_D \Delta \log \delta_1\, dA_0 \\ &= 2\pi - \int_D (G - \Delta \log \delta_1)\, dA_0,\end{aligned}$$

as desired. □

References

[do Carmo 1976] M. P. do Carmo, *Differential geometry of curves and surfaces*, Prentice-Hall, Englewood Cliffs, N.J., 1976. MR 52 #15253 Zbl 0326.53001

[Carroll et al. 2008] C. Carroll, A. Jacob, C. Quinn, and R. Walters, "The isoperimetric problem on planes with density", *Bull. Aust. Math. Soc.* **78**:2 (2008), 177–197. MR 2009i:53051 Zbl 1161.53049

[Corwin et al. 2006] I. Corwin, N. Hoffman, S. Hurder, V. Sesum, and Y. Xu, "Differential geometry of manifolds with density", *Rose-Hulman Und. Math. J.* **7**:1 (2006), article 2.

[Dubrovin et al. 1992] B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern geometry: methods and applications, I*, 2nd ed., Grad. Texts in Math. **93**, Springer, New York, 1992. MR 92h:53001 Zbl 0751.53001

[Howards et al. 1999] H. Howards, M. Hutchings, and F. Morgan, "The isoperimetric problem on surfaces", *Amer. Math. Monthly* **106**:5 (1999), 430–439. MR 2000i:52027 Zbl 1003.52011

[Morgan 1998] F. Morgan, *Riemannian geometry: a beginner's guide*, 2nd ed., A. K. Peters, Wellesley, MA, 1998. MR 98i:53001 Zbl 0911.53001

[Morgan 2009] F. Morgan, *Geometric measure theory: a beginner's guide*, 4th ed., Elsevier/Academic Press, Amsterdam, 2009. MR 2009i:49001 Zbl 1179.49050

Received: 2011-06-30 Revised: 2011-07-11 Accepted: 2011-07-11

ivan.corwin@gmail.com *Courant Institute of Mathematics, New York University, 251 Mercer Street, New York, NY 10012, United States*

Frank.Morgan@williams.edu *Department of Mathematics and Statistics, Williams College, 18 Hoxsey Street, Williamstown, MA 01267, United States*

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@mathscipub.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve
2011 vol. 4 no. 2